Capacity-Achieving Guessing Random Additive Noise Decoding

Ken R. Duffy, Jiange Li, and Muriel Médard, *Fellow, IEEE*

***Abstract*— We introduce a new algorithm for realizing maximum likelihood (ML) decoding for arbitrary codebooks in discrete channels with or without memory, in which the receiver rank-orders noise sequences from most likely to least likely. Subtracting noise from the received signal in that order, the first instance that results in a member of the codebook is the ML decoding. We name this algorithm GRAND for Guessing Random Additive Noise Decoding. We establish that GRAND is capacity-achieving when used with random codebooks. For rates below capacity, we identify error exponents, and for rates beyond capacity, we identify success exponents. We determine the scheme's complexity in terms of the number of computations that the receiver performs. For rates beyond capacity, this reveals thresholds for the number of guesses by which, if a member of the codebook is identified, that it is likely to be the transmitted code word. We introduce an approximate ML decoding scheme where the receiver abandons the search after a fixed number of queries, an approach we dub GRANDAB, for GRAND with ABandonment. While not an ML decoder, we establish that the algorithm GRANDAB is also capacity-achieving for an appropriate choice of abandonment threshold, and characterize its complexity, error, and success exponents. Worked examples are presented for Markovian noise that indicate these decoding schemes substantially outperform the brute force decoding approach.**

***Index Terms*— Discrete channels, maximum likelihood decoding, approximate ML decoding, error probability, channel coding.**

I. Introduction

CONSIDER a discrete channel with inputs, X^n, and outputs, Y^n, consisting of blocks of n symbols from a finite alphabet $\mathbb{A}$ of size $|\mathbb{A}|$. Assume that channel input is altered by random, not necessarily memoryless, noise, N^n, that is independent of the channel input and also takes values in $\mathbb{A}^n$. Assume that the function, $\oplus$, describing the channel's action,

$$Y^n = X^n \oplus N^n, \tag{1}$$

is invertible, so that knowing the output and input the noise can be recovered:

$$X^n = Y^n \ominus N^n. \tag{2}$$

Manuscript received April 5, 2018; revised January 21, 2019; accepted January 21, 2019. Date of publication January 31, 2019; date of current version June 14, 2019. This work was supported in part by the National Science Foundation under Grant 6932716. This paper was presented in part at the 2018 Information Theory and Applications Workshop and in part at the 2018 International Symposium on Information Theory, Colorado, USA.

K. R. Duffy is with the Hamilton Institute, Maynooth University, W23 F2H6 Maynooth, Ireland (e-mail: ken.duffy@mu.ie).

J. Li and M. Médard are with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: lijiange@mit.edu; medard@mit.edu).

Communicated by N. Merhav, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2019.2896110

To implement Maximum-Likelihood (ML) decoding, the sender and receiver first share a code-book $\mathcal{C}_n = \{c^{n,1}, \ldots, c^{n,M_n}\}$ consisting of M_n elements of $\mathbb{A}^n$. For a given channel output y^n, denote the conditional probability of the received sequence for each code-word in the code-book by

$$p(y^n|c^{n,i}) = P(y^n = c^{n,i} \oplus N^n) \quad \text{for } i \in \{1, \ldots, M_n\}. \tag{3}$$

The decoding produced by GRAND is then an element of the code-book that has the highest conditional likelihood of transmission given what was received,

$$\begin{aligned} c^{n,*} &\in \arg\max \left\{ p(y^n|c^{n,i}) : c^{n,i} \in \mathcal{C}_n \right\} \\ &= \arg\max \left\{ P(N^n = y^n \ominus c^{n,i}) : c^{n,i} \in \mathcal{C}_n \right\}, \end{aligned} \tag{4}$$

where we have used the invertibility of $\oplus$ for the final equality.

Code-book sizes are typically exponential in the block length n, $M_n \sim |\mathbb{A}|^{nR}$ and, taking logs throughout the article as base $|\mathbb{A}|$, we define the normalized rate of the code-book to be $R = \lim_n 1/n \log(M_n)$. Thus ML decoding would appear to be infeasible in practice for reasonable rates as it would seem that the receiver would have to either: A) perform $|\mathbb{A}|^{nR}$ conditional probability computations described in equation (3) followed by a rank ordering every time a signal is received; or B), in advance, perform $|\mathbb{A}|^{n(R+1)}$ computations described in equation (3), one for every $(c^{n,i}, y^n)$ pair, storing in a look-up table the resulting $|\mathbb{A}|^n$ ML decodings, one for each possible received sequence.

In the present paper we consider a distinct algorithm for ML decoding. The principle underlying the approach is for the receiver to rank-order noise sequences from most likely to least likely and then sequentially query whether the sequence that remains when the noise is removed from the received signal is an element of the code-book. For the channel structure described above, irrespective of how the code-book is constructed, the first instance where the answer is in the affirmative corresponds to the ML decoding. More formally, the receiver first creates an ordered list of noise sequences, $G : \mathbb{A}^n \mapsto \{1, \ldots, |\mathbb{A}|^n\}$, from most likely to least likely, breaking ties arbitrarily:

$$\begin{aligned} &G(z^{n,i}) \leq G(z^{n,j}) \\ &\text{if and only if } P(N^n = z^{n,i}) \geq P(N^n = z^{n,j}), \end{aligned} \tag{5}$$

where throughout this article lower case letters correspond to realizations of upper-case random variables, apart from for noise where z is used as n denotes the code block-length. For each received signal, the receiver executes the

TABLE I

DESCRIPTION OF ML DECODING BY GRAND. THE RECEIVER CREATES A RANK-ORDERED LIST OF NOISE FROM MOST LIKELY TO LEAST LIKELY BREAKING TIES ARBITRARILY, $z^{n,1}, z^{n,2}, \ldots$. IN THAT ORDER, GIVEN A RECEIVED SIGNAL y^n, THE RECEIVER SEQUENTIALLY SUBTRACTS THE NOISE $z^{n,i}$ AND QUERIES IF THE STRING THAT RESULTS, $y^n \ominus z^{n,i}$, IS AN ELEMENT OF THE CODE-BOOK, $\mathcal{C}_n$. THE FIRST STRING THAT IS IN THE CODE-BOOK, IS THE ML DECODING. IN THIS EXAMPLE, c^{n,i_1} IS THE FIRST ELEMENT OF THE CODE-BOOK TO BE IDENTIFIED, WHICH OCCURS ON THE THIRD NOISE GUESS. IN APPROXIMATE ML DECODING, GRANDAB, AFTER A FIXED NUMBER OF QUERIES THE RECEIVER ABANDONS THE QUESTIONING AND DECLARES AN ERROR

Noise guessing order	1	2	3	4	5	6	...
Noise from most likely to least likely	$z^{n,1}$	$z^{n,2}$	$z^{n,3}$	$z^{n,4}$	$z^{n,5}$	$z^{n,6}$	...
String queried for membership of the code-book $\mathscr{C}_n$	$y^n \ominus z^{n,1}$	$y^n \ominus z^{n,2}$	$y^n \ominus z^{n,3}$	$y^n \ominus z^{n,4}$	$y^n \ominus z^{n,5}$	$y^n \ominus z^{n,6}$	...
Location of code-book elements			c^{n,i_1}		c^{n,i_2}		...

following algorithm, which we call GRAND for Guessing Random Additive Noised Decoding:

- Given channel output y^n, initialize $i = 1$ and set z^n to be the most likely noise sequence, i.e. the z^n such that $G(z^n) = i$.
- While $x^n = y^n \ominus z^n \notin \mathcal{C}_n$, increase i by 1 and set z^n to be the next most likely noise sequence, i.e. the z^n such that $G(z^n) = i$.
- The x^n that results from this while loop is the decoded element.

An example of this process is described in Table I.

To see that GRAND corresponds to ML decoding for channels of the sort described in equations (1) and (2), note that, owing to the definition of $c^{n,*}$ in equation (4),

$$P(N^n = y^n \ominus c^{n,*}) \geq P(N^n = y^n \ominus c^{n,i}) \text{ for all } c^{n,i} \in \mathcal{C}_n.$$

Thus the scheme does, indeed, identify an ML decoding. The premise of the present paper is that this new scheme, GRAND, has a complexity that decreases as code-book rate increases even though the more direct approach described in equation (4) sees steeply increasing complexity.

In Section III-A, the performance of the algorithm is established in terms of its maximum achievable rate, which is a property of ML decoding rather than being particular to the present GRAND scheme, and the number of computations the receiver must perform until decoding, which is dependent on the scheme. With some mild ergodicity conditions imposed on the noise process, we prove that GRAND is capacity achieving with uniform-at-random code-books. We determine asymptotic error exponents, as well as providing success exponents for rates above capacity. We identify the asymptotic complexity of GRAND in terms of the number of local operations the receiver must perform per received block in executing the algorithm.

Based on this new noise-centric design ethos for ML decoding and the intuition that comes from its analysis, we introduce a new approximate ML decoder in Section III-B, an approach we dub GRANDAB for GRAND with ABandonment. In this variant of GRAND, the receiver abandons identification of the transmitted code word if no element of the code-book is identified after a pre-defined number of noise removal queries. GRANDAB is not a ML decoder, as the algorithm sometimes terminates without returning an element of the code-book. Despite that, we establish that GRANDAB is also capacity achieving for random code-books once the abandonment threshold is set for after all elements of the Shannon Typical Set of the noise are queried, and we determine the exponent for the likelihood of abandonment. By abandoning after a fixed number of queries, an upper-bound on complexity is ensured.

To determine these algorithmic properties, we leverage recent results in the study of guesswork. We recall one theorem from the literature and establish several new ones. As they may appear somewhat mathematically involved, we begin by explaining the intuitive meaning behind them.

Reference [1, Th. 1] provides a Large Deviation Principle (LDP) as the block length, n, becomes large, for the distribution of the logarithm of the number of guesses needed until the actual noise in the channel is queried, $G(N^n)$. On its own, this result provides us with an upper-bound on the complexity of the scheme, but it can be augmented in the case of uniformly selected code-books. That is, where the input elements X^n in equation (1) are chosen uniformly at random from a code-book $\mathcal{C}_n$ that itself consists of a collection of uniformly selected elements of $\mathbb{A}^n$.

Theorem 2 is new and establishes properties of the number of guesses that would be made until an element of the code-book that was not the channel input is identified. Here we leverage the fact that for uniformly distributed code-books the location of each of these elements in the guessing order outlined in Table I are uniform in $\{1, \ldots, |\mathbb{A}|^n\}$. As a result, the distribution of the number of guesses until any non-input element of the code-book is hit upon is distributed as the minimum of M_n such uniform random variables. When $M_n \approx |\mathbb{A}|^{nR}$ and n becomes large, the resulting minimum is essentially the discretization of an exponential distribution with rate $|\mathbb{A}|^{-n(1-R)}$ so that the receiver will identify a code-word in, on average, approximately $|\mathbb{A}|^{n(1-R)}$ guesses. Note, in particular, that as R increases and the code-book becomes more dense and efficient, while the number of computations in the brute-force approach to ML decoding increases, the noise guessing approach experiences the reverse phenomenon.

The ML decoding algorithm introduced in the present paper is essentially a race between the two guessing processes mentioned above. If the number of guesses required to identify the true noise is less than the number of guesses to identify any other element of the code-book, then GRAND provides the correct answer on termination. Combining the two earlier results in two different ways first recovers the Channel Coding Theorem as Proposition 1 via this new guessing argument. Namely, with R being the normalized code-book rate, H being the normalized Shannon entropy rate of the noise base $|\mathbb{A}|$, and with $1 - H$ being the channel capacity, so long as $R < 1 - H$ then the ML decoder will correctly identify the input for long enough blocks. The guessing argument provides

asymptotic exponents for the probability that the ML decoding is an error if the code-book is within capacity, as well as for the probability that the ML decoding is correct if the code-book rate is beyond capacity. Both of these error and success exponents are convex functions of the code-book rate near capacity and approach zero at capacity, hinting at smooth degradation in performance near capacity.

Combining Theorems 1 and 2 in a distinct fashion akin to that used in [2] to study multi-user guesswork, Proposition 2 characterizes the complexity of the scheme in terms of the distribution of the number of guesses to termination. This approach allows us to determine some subtle performance features of the scheme when code-books rates are beyond capacity. Theorem 3 establishes that the circumstances beyond capacity under which the ML decoding is likely to be correct decoding should the noise guessing complete quickly. In particular, this phenomenon occurs if the code-book rate is less than one minus the min-entropy rate of the noise.

Interpreting the results of Propositions 1 and 2 in light of the noise guessing algorithm leads us to propose a new approximate ML decoder, GRANDAB. In GRANDAB, if no code-book element is identified by the receiver after $|\mathbb{A}|^{n(H+\delta)}$ queries for some $\delta > 0$, the receiver abandons guessing and decoding results in an error. While it is not an ML decoder, we prove in Proposition 3 that GRANDAB is also capacity-achieving for any $\delta > 0$. Thus GRANDAB has the capacity achieving qualities of ML decoding with a guaranteed upper bound on the number of computations performed by the receiver. This can result in a significant saving over GRAND in terms of complexity as the average number of queries required to identify the true noise in the system grows with an exponent of Rényi entropy rate $1/2$.

In Section IV the performance of GRAND and GRANDAB are illustrated for bursty Markovian channel noise as, crucially, all of the results in this paper hold for channels with memory, a point we investigate in Section V. For memoryless channels, however, the guessing approach enables finer approximations to the computation of block error probabilities than asymptotic exponents and these are used for the BSC in Section IV-C. In Section V we conclude with a discussion of implementation and further potential of the principles underlying the decoding algorithms introduced here.

II. Related Work

Large deviation style arguments that are employed to establish error exponents in both source and channel coding are typically variants of Sanov's Theorem [3] and the method of types. If sources are assumed to have properties such as being independently and identically distributed (IID) or Markovian, identification of non-asymptotic pre-factors can be possible. For error exponents in source coding, these methods have been used extensively, originally for asymptotically error-free source coding with IID and Markov sources [4]–[6], and then for variable-length and lossy source coding of IID and stationary sources [7]. For channel coding of Discrete Memoryless Channels (DMCs), error exponents were first identified by Gallager [8] by direct arguments. In unpublished notes that are available on the web, Montanari and Forney [9] provide a relationship between Gallager's error exponent and the exponent obtained through large deviation considerations of channel coding arguments using asymptotic equipartition principles. More recently, an approach along these lines has been used to study joint source-channel coding [10]. As an aside, we remark that an alternate means of establishing the results in [10] would have been be to combine the results of [7] with the generalization in [11] of [12] and [13] using a method of types.

While the arguments used in the papers referenced above are essentially based on variants and refinements of the Large Deviation Principle (LDP) for empirical measures, we instead analyze our proposed approach starting from a completely distinct angle: the recently established LDP for Massey's guesswork [14]. That LDP is a development from earlier results that identify scaling exponents for moments of guesswork in terms of Rényi entropy rates [15]–[17]. Given the explicit relationship between the guesswork process and the noise guessing approach, this seems the most natural line of attack. In [18] Arikan establishes LDP bounds for conditional probability rank. The full large deviation principle, which we employ here, is proven in [1].

The connection between source coding and guesswork was first noted by Arikan and Merhav [19], and has been established by Hanawal and Sundaresan [20]. For channel coding, a connection between guesswork and error exponent analysis was proved by Arikan for sequential decoding of tree codes [15], such as classic convolutional codes [21]. Sequential decoding, introduced by Wozencraft [22] and Wozencraft and Reiffen [23], is a variant of ML decoding for tree codes. To ensure low computational complexity of sequential decoding of convolutional codes, rates are generally kept below a computational cutoff rate [15], [22], [24]–[29]. A survey of the historical rationale for cut-off rate design can be found in [30]. Several schemes have sought to improve the cut-off rate, including Pinsker's concatenated code with an inner block code and outer sequential decoder [31], as well as Massey's "splitting" argument for quaternary erasure channel [32]. A general framework for designing codes that increase the cutoff rate is discussed in [33]. Polar Coding, which is capacity achieving for binary DMCs [34], fits into that framework.

For linear block codes, an ML decoding method exists that has complexity bounded by $2^{n(1-R)}$ (in the current article's notation) [35]. As the complexity of brute force ML decoding is 2^{nR}, that approach is preferable when $1 - R < R$, that is when $R > 1/2$. For rates below capacity, $R < 1-H$ and hence $H < 1 - R$. GRANDAB's complexity $2^{n(H+\delta)}$, for arbitrary $\delta > 0$, is thus lower than the one provided by [35], except for low code rates where the complexity of brute force ML decoding is preferable. The approach taken in [35] is based on a trellis decoding method for linear convolutional codes akin to the one independently derived in [36], in which terminated, or so called blocked, convolutional codes are also considered.

To formally establish capacity and complexity results, in the present work, we do not envisage designing codes, but using random ones. For the channels we are considering, Shannon's [37] uniform random code-book plus ML decoding argument affords capacity, but for codes of sufficient length

that approach capacity, decoding methods for random codes are prohibitively complex with existing methods, as explained in the introduction. The core performance idea here is to leverage the fact that the noise is typically highly non-uniform, rendering its identification through guessing less onerous than performing a computation for every element of the code-book.

While our model employs uniformly distributed code-words, we analyze substantially more general noise processes than the DMC. For the DMC, the error exponent we derive necessarily matches Gallager's. That is unsurprising, as he proves it was tight for the average code [38], and this fact has recently generalized to random linear codes [39] for channels for which uniform code-books are optimal. As an aside, we remark that the result in [38] is echoed in the source coding domain in [7], which shows, via asymptotic equipartition style arguments, that almost all random code-books provide in effect the same compression performance. Thus, one might suspect that results analogous to those in [39] are likely to hold also for source coding [40] and network coding [41], [42].

The mathematical approach we take naturally lends itself to the determination of decay exponents in the probability of success when coding above capacity. The question of success for codes operating above capacity is a long-standing, though perhaps less well studied than that of errors below capacity [43]–[45]. For a DMC, lower bounds [46] that are coincident with upper bounds [47] are known to exist. Here, the derivation of these exponents come hand-in-hand with the determination of error exponents, and hold for the same broad class of noise processes.

GRAND employs ordered statistics of noise for decoding, but the code-book is only used when checking if a proposed decoded code word pertains to the code-book. The noise statistics may be obtained by arbitrary means and are not dependent on examining the decoder's output. This approach differs from Ordered Statistics Decoding (OSD) [48], [49], which uses the statistics derived from syndrome computations to update soft information in decoding linear bock codes, or from Turbo-style systems that blend decoding with soft information, see for instance [50]–[53].

As ML decoding is generally too onerous from a complexity perspective, the use of approximate ML decoding is, under different guises, almost omnipresent in decoding algorithms. The approach GRANDAB takes, that of stopping after a given set of guesses, is redolent of limited search approaches commonly used in the decoding of convolutional codes, such as reduced state sequence estimation (RSSE) and related techniques that limit the search space in sequential decoding [54]–[61]. This latter family of techniques uses the received sequence as a starting point, rather than consider the noise itself as we do in GRANDAB, and most have not been formally established to be capacity achieving.

III. Analysis

A. ML Decoding by Guesswork

We begin with the assumption we shall make on the noise process. Recall that log is taken base $|\mathbb{A}|$ throughout.

Assuming it exists, define the Rényi entropy rate of the noise $\{N^n\}$ process with parameter $\alpha \in (0,1) \cup (1,\infty)$ to be

$$H_\alpha = \lim_{n\to\infty} \frac{1}{n} \frac{1}{1-\alpha} \log \left(\sum_{z^n \in \mathbb{A}^n} P(N^n = z^n)^\alpha \right),$$

with $H = H_1$ being the Shannon entropy rate of the noise. Denote the min-entropy rate of the noise by $H_{\min} = \lim_{\alpha\to\infty} H_\alpha$.

Assumption 1. *We assume that*

$$\Lambda^N(\alpha) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}(G(N^n)^\alpha) = \begin{cases} \alpha H_{1/(1+\alpha)} & \text{for } \alpha \in (-1,\infty) \\ -H_{\min} & \text{for } \alpha \le -1, \end{cases} \tag{6}$$

and that the derivative of $\Lambda^N(\alpha)$ is continuous on the range $\alpha \in (-1,\infty)$.

Assumption 1 is known to be satisfied for a broad range of sources including i.i.d. [15], Markovian [16], a large class of general, stationary processes [17] and others [62]; the condition for $\alpha \le -1$ is established for all of these in [1].

Note that, by setting $\alpha = 1$, as first identified by Arikan [15], from equation (6) one has that the average number of guesses required to identify the true noise grows exponentially in block size, n, with Rényi entropy rate at parameter $1/2$, $H_{1/2}$, which is no smaller than the Shannon entropy rate, H, of the noise. Note also that $\Lambda^N(\alpha)$ has a continuous derivative everywhere except potentially at $\alpha = -1$. An operational meaning to the discontinuous derivative when evaluated from above is identified in [1], where the value of the discontinuity captures the exponential growth in n of the size of the set of most-likely noise sequences.

Example. *For a BSC with $\mathbb{A} = \{0,1\}$ and an additive channel mod 2, and $P(N^1 = 1) = p$,*

$$\Lambda^N(\alpha) = \begin{cases} (1+\alpha)\log\left((1-p)^{\frac{1}{1+\alpha}} + p^{\frac{1}{1+\alpha}}\right) & \text{if } \alpha \in (-1,\infty) \\ \log(\max(1-p,p)) & \text{if } \alpha \le -1. \end{cases} \tag{7}$$

Plots of $\Lambda^N(\alpha)$ can be found in Fig. 1.

From equation (6), Λ^N can be identified as the scaled cumulant generating function for the process $\{1/n \log G(N^n)\}$ [3] and so Λ^N is necessarily convex. Moreover, that identification suggested that this process may satisfy a Large Deviation Principle (LDP) [19], [63], which is proved in [1] and used in [64]–[67].

Theorem 1. *(LDP for Guessing the Noise [1]). Under Assumption 1, $\{1/n \log G(N^n)\}$ satisfies the Large Deviation Principle with the convex lower-semi continuous rate function, $I^N : [0,1] \to [0,\infty]$,*

$$I^N(x) := \sup_{\alpha\in\mathbb{R}} \{x\alpha - \Lambda^N(\alpha)\}, \tag{8}$$

which is the Legendre-Fenchel transform of Λ^N.

In particular: $I^N(0) = H_{\min}$, the min-entropy rate of the noise; $I^N(x)$ is linear on $[0,\gamma]$, where $\gamma := \lim_{\alpha\downarrow -1} d/d\alpha\, \Lambda^N(\alpha)$, and then strictly convex thereafter while finite; and $I^N(x) = 0$ if and only if $x = H$, the Shannon entropy rate of the noise.

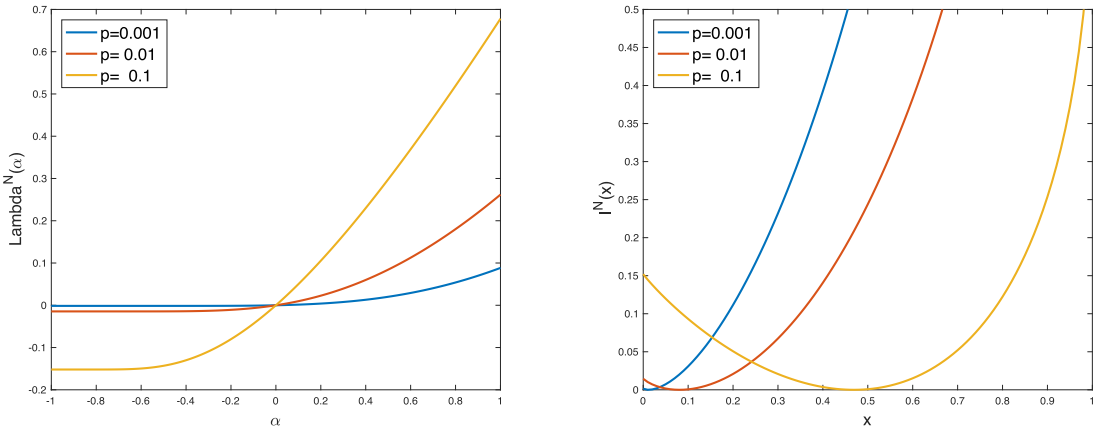

Fig. 1. Guesswork rate function. Example: $\mathbb{A} = \{0, 1\}$, BSC channel, noise N^n made of i.i.d. Bernoulli symbols with $P(N^1 = 1) = p \in (0, 1)$. Left panel: scaled cumulant generating function, Λ^N, for the noise process $\{1/n \log G(N^n)\}$ as determined by the explicit expression in (7). Right panel: rate function for the same process, I^N, defined in equation (8) and determined numerically. Roughly speaking $\log P(n^{-1} \log G(N^n) \approx x) \approx -nI^N(x)$. Note that $I^N(0) = H_{\min}$, the min-entropy of the noise, and that the rate function is zero at the Shannon entropy of the noise, $I^N(H) = 0$.

Intuitively, this result says that, for fixed (a, b), as n goes to infinity

$$\log P\left(\frac{1}{n}\log G(N^n) \in (a, b)\right) \approx -n \inf_{x\in(a,b)} I^N(x)$$

for large n. As well as providing this approximation, one of the primary advantages of a LDP over knowing how moments scale from Λ^N is that it is covariant in the sense that LDPs are preserved by continuous maps [3, Th. 4.2.1], and we shall repeatedly use that property to combine distinct LDPs.

Example. *While there is no closed form expression for I^N for the BSC, it can be readily computed numerically and examples are provided in Fig. 1.*

For random code-books, the second theorem provides a LDP for the number of guesses on the noise that will be made until identifying an element of the code-book that is not the transmitted code-word. The key realization is that, if elements of the code-book have been selected uniformly at random, the location of the non-transmitted code-book elements in the ordered list of noise guesses are also uniform. Let $U^{n,1}, \ldots, U^{n,M_n}$ be independent random variables, each uniformly distributed in $\{1, \ldots, |\mathbb{A}|^n\}$ and define

$$U^n = \min_i U^{n,i}.$$

Assumption 2. *Assume that $\lim_{n\to\infty} n^{-1} \log M_n = R$ for some $R > 0$.*

Theorem 2. *(LDP for Guessing a Non-Transmitted Code-Word). Under Assumption 2, as n becomes large, U^n is approximately exponentially distributed with rate $|\mathbb{A}|^{-n(1-R)}$,*

$$\lim_{n\to\infty} P(|\mathbb{A}|^{-n(1-R)} U^n > x) = e^{-x} \text{ for all } x > 0. \quad (9)$$

Moreover, $\{1/n \log U^n\}$ satisfies the large deviation principle with lower semi-continuous rate function

$$I^U(x) = \begin{cases} 1 - R - x & \text{if } x \in [0, 1 - R] \\ +\infty & \text{otherwise} \end{cases} \quad (10)$$

and

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}(U^n) = 1 - R.$$

Proof: We begin by observing that

$$P\left(U^n > |\mathbb{A}|^{xn}\right) = \prod_{i=1}^{M_n} P\left(U^{n,i} > |\mathbb{A}|^{xn}\right) = \left(1 - \frac{\lceil |\mathbb{A}|^{xn} \rceil}{|\mathbb{A}|^n}\right)^{M_n}.$$

Setting $x = 1 - R$ and making use of L'Hospital's rule, as $\lim_{n\to\infty} n^{-1} \log M_n = R$ we have that for $y > 0$

$$\lim_{n\to\infty} P(|\mathbb{A}|^{-n(1-R)} U^n > y) = \lim_{n\to\infty} \left(1 - y|\mathbb{A}|^{-nR}\right)^{|\mathbb{A}|^{nR}} = e^{-y},$$

giving equation (9).

As $[0, 1]$ is compact, in order to establish the LDP it is sufficient [3] to prove that

$$\begin{aligned} &\lim_{\epsilon\downarrow 0} \liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\log U^n \in (x - \epsilon, x + \epsilon)\right) \\ &= \lim_{\epsilon\downarrow 0} \limsup_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\log U^n \in (x - \epsilon, x + \epsilon)\right) \\ &= -I^U(x) \end{aligned} \quad (11)$$

for all $x \in [0, 1]$. Using the earlier observation, we have the following limiting equality for the survival function

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\log U^n > x\right) &= \lim_{n\to\infty} \frac{M_n}{n} \log\left(1 - \frac{\lceil |\mathbb{A}|^{xn} \rceil}{|\mathbb{A}|^n}\right) \\ &= \lim_{n\to\infty} \frac{|\mathbb{A}|^{nR}}{n} \log\left(1 - |\mathbb{A}|^{n(x-1)}\right) \\ &= -\lim_{n\to\infty} \frac{1}{n} |\mathbb{A}|^{n(R+x-1)} \\ &= \begin{cases} 0 & \text{if } x \in [0, 1 - R] \\ -\infty & \text{if } x \in (1 - R, 1). \end{cases} \end{aligned}$$

From this, we can confirm the veracity of equation (11) for all $x \in (1 - R, 1]$:

$$\begin{aligned} &\lim_{\epsilon\downarrow 0} \lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\log U^n \in (x - \epsilon, x + \epsilon)\right) \\ &\le \lim_{\epsilon\downarrow 0} \lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\log U^n > x - \epsilon\right) = -\infty. \end{aligned}$$

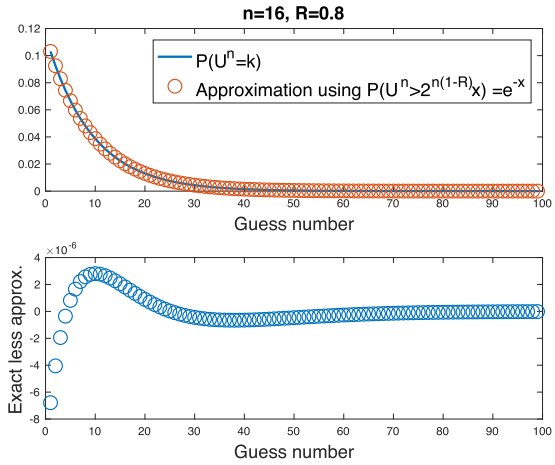

Fig. 2. Example: $\mathbb{A} = \{0, 1\}$, block length $n = 16$ and $R = 4/5$. Upper panel: compares exact computation of $P(U^n = k)$ (blue line) with the exponential distribution approximation given in equation (9) (orange circles) for first 100 guesses. Lower panel: the difference between the exact and approximate values.

The corresponding equality for the cumulative distribution function can be obtained by first noting that, by the Binomial theorem,

$$\lim_{n\to\infty} \frac{\left(1 - |\mathbb{A}|^{n(x-1)}\right)^{|\mathbb{A}|^{nR}}}{1 - |\mathbb{A}|^{n(R+x-1)}} = 1 \text{ if } x \in [0, 1 - R),$$

while if $x = 1 - R$ the limit of the numerator in the above equation is $\exp(-1)$. Thus to prove that equation (11) holds for $x \in [0, 1 - R]$, we have

$$\begin{aligned}
&\lim_{\epsilon\downarrow 0}\lim_{n\to\infty}\frac{1}{n}\log P\left(\frac{1}{n}\log U^n \in (x-\epsilon, x+\epsilon)\right)\\
&= \lim_{\epsilon\downarrow 0}\lim_{n\to\infty}\frac{1}{n}\log\left(P\left(\frac{1}{n}\log U^n < x+\epsilon\right) - P\left(\frac{1}{n}\log U^n \le x-\epsilon\right)\right)\\
&= \lim_{\epsilon\downarrow 0}\lim_{n\to\infty}\frac{1}{n}\log\left(\left(1-\left(1-\frac{\lceil|\mathbb{A}|^{(x+\epsilon)n}\rceil}{|\mathbb{A}|^n}\right)^{M_n}\right)\right.\\
&\qquad\left.-\left(1-\left(1-\frac{\lceil|\mathbb{A}|^{(x-\epsilon)n}\rceil}{|\mathbb{A}|^n}\right)^{M_n}\right)\right)\\
&= \lim_{\epsilon\downarrow 0}\lim_{n\to\infty}\frac{1}{n}\log\left(|\mathbb{A}|^{n(\min(R+x+\epsilon-1,0))} - |\mathbb{A}|^{n(R+x-\epsilon-1)}\right)\\
&= -(1-R-x),
\end{aligned}$$

as $R + x - \epsilon - 1 < 0$ for $x \in [0, 1 - R]$.

The scaling result for the mean of U^n follows from the application of Varadhan's Theorem [3, Th. 4.3.1], giving

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}(U^n) = \sup_{x\in[0,1-R]}\left(x - I^U(x)\right) = 1 - R. \quad \blacksquare$$

Equation (9) provides a highly accurate approximation of the distribution of U^n, that it is essentially exponentially distributed with rate $|\mathbb{A}|^{-n(1-R)}$ giving rise to a mean of $|\mathbb{A}|^{n(1-R)}$. This is illustrated in Fig. 2 for a block length of $n = 16$ and a code-book of rate $R = 4/5$, and becomes more precise as n increases. We will use this approximation to make near exact computations of block error probabilities and complexity for the BSC in Section IV-C. To establish the general channel coding and complexity results, however, it is the LDP that is needed. On the scale of large deviations, Theorem 2 effectively says that, for large n, the first non-transmitted code-word will be encountered in no more than order $|\mathbb{A}|^{n(1-R)}$ guesses.

Combining Theorems 1 and 2 enables us to provide a guessing based proof of Channel Coding Theorem. Recalling that logarithms are taken base $|\mathbb{A}|$, let h denote the Shannon entropy of a random variable and let I denote mutual information. For channels introduced in equations (1) and (2), capacity is upper bounded by $1 - H$ as follows:

$$C \le \limsup_{n\to\infty}\frac{1}{n}\sup I(X^n; Y^n) \le 1 - \lim_{n\to\infty}\frac{h(N^n)}{n} = 1 - H,$$

where we have upper-bounded the entropy rate of the input, $h(X^n)$, by its maximum, n, and used the fact that the channel is invertible [i.e. equation (2)], while the entropy rate of the noise exists, owing to Assumption 1. The proposition that follows establishes, through the use of a uniform-at-random code-book and GRAND, that this upper bound is achieved for all noise processes satisfying Assumption 1. We define the success rate

$$s(R) = -\lim_{n\to\infty}\frac{1}{n}\log P(U^n \ge G(N^n)),$$

which is the decay rate in the probability of correct decoding, and evaluate it in the case where the code rate exceeds capacity.

Proposition 1 *(Channel Coding Theorem With GRAND). Under Assumptions 1 and 2, with I^U defined in equation* (10) *and I^N in equation* (8)*, we have the following.*

1) If the code-book rate is less than the capacity, $R < 1-H$, then

$$\lim_{n\to\infty}\frac{1}{n}\log P(U^n \le G(N^n)) = -\inf_{a\in[H,1-R]}\{I^U(a) + I^N(a)\} < 0,$$

so that the probability that GRAND does not correctly identify the transmitted code-word decays exponentially in the block length n. If, in addition, x^* exists such that

$$\frac{d}{dx}I^N(x)|_{x=x^*} = 1, \tag{12}$$

then the error rate simplifies further to

$$\begin{aligned}
\epsilon(R) &= -\lim_{n\to\infty}\frac{1}{n}\log P(U^n \le G(N^n))\\
&= \begin{cases} 1 - R - H_{1/2} & \text{if } R \in (0, 1 - x^*)\\ I^N(1-R) & \text{if } R \in [1-x^*, 1-H). \end{cases}
\end{aligned} \tag{13}$$

Moreover,

$$s(R) = \lim_{n\to\infty}\frac{1}{n}\log P(U^n \ge G(N^n)) = 0$$

so that the probability that GRAND does not provide the true channel does not decay exponentially in n.

2) If, instead, the code-book rate is greater than the capacity, $R > 1-H$, then the probability of an error is not decaying exponentially in n,

$$\lim_{n\to\infty}\frac{1}{n}\log P(U^n \le G(N^n)) = 0.$$

However,

$$s(R) = I^N(1-R), \tag{14}$$

is strictly positive, so that the probability that decoding produced by GRAND is the transmitted code-word does decay exponentially in n.

Proof: As $\{1/n \log G(N^n)\}$ and $\{1/n \log U^n\}$ are independent processes, $\{(1/n \log G(N^n), 1/n \log U^n)\}$ satisfies the LDP with rate function $I^N(x) + I^U(y)$. The LDP for $\{1/n \log U^n/G(N^n)\}$ then follows from an application of contraction principle, [3, Th. 4.2.1], with the continuous function $f(x, y) = x - y$, giving

$$\begin{aligned} I^{U/N}(x) &= \inf_{a,b} \left\{ I^N(a) + I^U(b) : f(a,b) = a - b = x \right\} \\ &= \inf_{a \in [0,1-R]} \{ I^U(a) + I^N(a-x) \}. \end{aligned}$$

Noting the following equality

$$P(U^n \le G(N^n)) = P\left(\frac{1}{n} \log \frac{U^n}{G(N^n)} \le 0 \right),$$

we can use the LDP for $\{1/n \log U^n/G(N^n)\}$ to determine asymptotics for the likelihood that fewer queries are necessary to determine a non-transmitted element of the code-book than the truly transmitted element. From the LDP lower and upper bounds,

$$\begin{aligned} -\inf_{x<0} I^{U/N}(x) &\le \liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} \log \frac{U^n}{G(N^n)} \le 0 \right) \\ &\le \limsup_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} \log \frac{U^n}{G(N^n)} \le 0 \right) \\ &\le -\inf_{x\le 0} I^{U/N}(x). \end{aligned}$$

For the limit to exist, we require that $\inf_{x<0} I^{U/N}(x) = \inf_{x\le 0} I^{U/N}(x)$. Consider $I^{U/N}(0) = \inf_{a\in[0,1-R]}\{I^U(a) + I^N(a)\} = I^U(a^*) + I^N(a^*) < \infty$, where a^* necessarily exists as I^U and I^N are lower-semicontinuous. As we have assumed $H > 0$, $a^* > 0$ and $I^U(a^*) + I^N(a^*)$ is then arbitrarily well approximated by $I^U(a^*) + I^N(a^* - \epsilon)$ as I^N is continuous where it is finite, so the above limit exits. The following simplification is achieved by changing the order the infima are taken in:

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log P(U^n \le G(N^n)) &= -\inf_{x\le 0} I^{U/N}(x) \\ &= -\inf_{x\le 0} \inf_{a\in[0,1-R]} \{I^U(a) + I^N(a-x)\} \\ &= -\inf_{a\in[0,1-R]} \{I^U(a) + \inf_{y\le a} I^N(y)\}. \end{aligned} \tag{15}$$

Starting from

$$P(U^n \ge G(N^n)) = P\left(\tfrac{1}{n} \log \tfrac{U^n}{G(N^n)} \ge 0 \right),$$

similar logic, but with an additional simplification due to the form of I^U found equation (10), leads to

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log P(U^n \ge G(N^n)) &= -\inf_{x\ge 0} I^{U/N}(x) \\ &= -\inf_{x\ge 0} \inf_{a\in[0,1-R]} \{I^U(a) + I^N(a-x)\} \\ &= -\inf_{a\in[0,1-R]} \{I^U(a) + \inf_{y\le a} I^N(y)\} \\ &= -\inf_{x\in[0,1-R]} I^N(x). \end{aligned} \tag{16}$$

(a) For the within-capacity result, if $R < 1 - H$, then $H < 1 - R$. Considering the right hand side of equation (15) as both I^U and I^N are decreasing on $[0, H]$ and I^N is either infinite or increasing on $[H, 1 - R]$,

$$\inf_{a\in[0,1-R]} \{I^U(a) + \inf_{y\ge a} I^N(y)\} = \inf_{a\in[H,1-R]} \{I^U(a) + I^N(a)\}.$$

This quantity is strictly positive, as I^U is strictly decreasing to zero on $[H, 1 - R]$, while I^N is strictly increasing from zero on the same range. To get the additional simplification to equation (13), note that, as I^N is strictly convex to the right of H, I^U is decreasing at rate 1 and x^* is defined to be the value at which I^N is increasing with rate 1, then $\inf_{a\in[H,1-R]}\{I^U(a) + I^N(a)\}$ is either $I^N(1-R)$ if $x^* > 1-R$ or $I^U(x^*) + I^N(x^*)$. Now $I^N(x^*) = x^* - H_{1/2}$, so that $I^U(x^*) + I^N(x^*) = 1 - R - x^* + x^* - H_{1/2}$ and the result follows. On the other hand,

$$\inf_{x\in[0,1-R]} I^N(x) = I^N(H) = 0$$

and so the right hand side of equation (16) is zero.

(b) For the beyond-capacity result if, alternatively, $R > 1 - H$, then $H > 1 - R$ and

$$\inf_{a\in[0,1-R]} \{I^U(a) + \inf_{y\ge a} I^N(y)\} = I^U(1-R) + I^N(H) = 0,$$

and so the right hand side of equation (15) is zero. While

$$\inf_{x\in[0,1-R]} I^N(x) = I^N(1-R) > 0,$$

so that the right hand side of (16) is strictly greater than zero. ∎

Proposition 1 not only proves the Channel Coding Theorem, but also provides exact asymptotic error exponents when the rate of the code-book, R, is within capacity, $1 - H$, and success exponents for when the rate is beyond capacity. For memoryless channels, the error rate in equation (13) coincides with that in [8, Th. 2], where the linear followed by strictly convex phenomenon was first identified. Proposition 1 establishes that phenomenon for more general noise processes.

The point $1 - x^*$ in equation (13), where the error exponent goes from being linear in the code-book rate to strictly convex in equation (13), is dubbed the critical rate by Gallager for memoryless channels and can be given a simple interpretation in terms of the noise guessing GRAND undertakes for general noise processes. For code-book rates R beyond the critical rate, in the large n limit an error occurs because the uniform code-book is typical, but the noise is exceptionally unlikely and far down the guessing order. For code-book rates below the critical rate, it requires an average number of guesses to

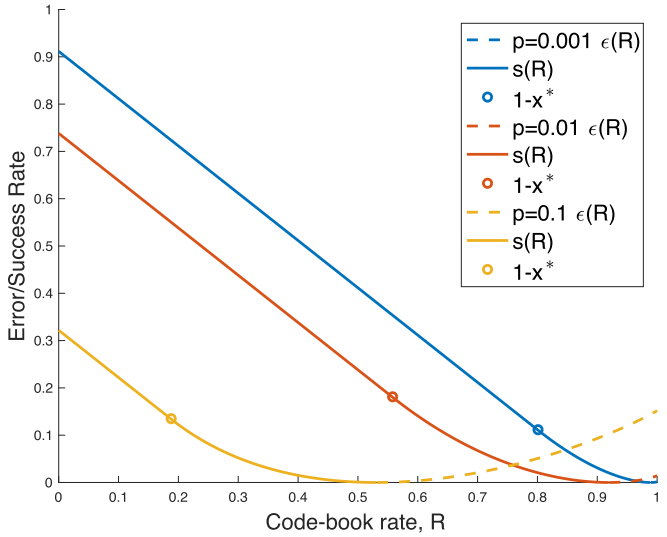

Fig. 3. GRAND decoding error and success exponents. Example: $\mathbb{A} = \{0, 1\}$, BSC channel, noise N^n made of i.i.d. Bernoulli symbols with $P(N^1 = 1) = p \in (0, 1)$. Code-book consisting of $M_n \approx 2^{nR}$ code-words, uniformly selected in $\mathbb{A}^n$. When the code-book rate, R, is less than channel capacity, $1 - H$, the probability that a code-word that was not sent is encountered during noise guessing before the transmitted code-word, $P(U^n < G(N^n))$, decays exponentially in block length n with rate $\epsilon(R)$ given by the solid line as determined by equation (13), which coincides in this case with Gallager's error exponent. The point $1-x^*$ marks the critical rate where the error-rate changes from linear to strictly convex. For code-books rates that are beyond capacity, $R > 1 - H$, the probability that the transmitted code-word is identified before a non-transmitted code-word, $P(G(N^n) < U^n)$, decays exponentially in n with rate $s(R)$ from equation (14), indicated by the dashed line.

identify the true noise, which is why the Rényi entropy rate with parameter $1/2$ appears, but the uniform code-book has an unusually early entry in the noise-guessing ordered list, resulting in an error.

Proposition 1 also provides success exponents for rates above capacity. Here the interpretation of the success rate in equation (14) is that, if the code-book rate R is too high for capacity, $1-H$, in the large n limit a successful decoding will occur if the non-transmitted code-book elements are typically distributed, but the noise is unusually highly likely, such that it is identified first, just prior to a non-transmitted element of the code-book.

Example. *For the BSC, example plots of these curves are provided in Fig. 3. Note that as I^N is a convex function that is zero at H, the error and success exponents are both smooth, near-zero functions around capacity, $R = 1-H$. This suggests that GRAND experiences graceful degradation in performance near capacity.*

We can combine Theorems 1 and 2 in a distinct way to determine the asymptotic complexity of the new ML decoding scheme in terms of the number of guesses until an ML decoding, correct or incorrect, is identified:

$$D^n := \min(G(N^n), U^n). \tag{17}$$

That is, GRAND terminates at either identification of the noise that was in the channel or when a non-transmitted element of the code-book is unintentionally identified, whichever occurs first. On the scale of large deviations, if the code-book is within capacity, $R < 1 - H$, then the sole impact of the code-book is to curtail excessive guessing when unusual noise occurs.

Proposition 2 *(Guessing Complexity of GRAND). Under Assumptions 1 and 2, $\{1/n \log D^n\}$ satisfies a LDP with a lower-semicontinuous rate function, I^D.*

1) If $R < 1-H$, then the input code-word will be recovered in the large deviations limit with unaffected likelihoods, and the impact of the code-book is to curtail guessing of unlikely inputs:

$$I^D(x) = \begin{cases} I^N(x) & \text{if } x \in [0, 1-R] \\ +\infty & \text{if } x > 1-R. \end{cases} \tag{18}$$

The average number of guesses until GRAND finds a decoding satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log E(D^n) = \min\left(H_{1/2}, 1-R\right).$$

2) If $R > 1-H$, the code-book rate is higher than capacity and

$$I^D(x) = \begin{cases} \min\left(I^N(x), I^U(x)\right) & \text{if } x \in [0, 1-R] \\ +\infty & \text{if } x > 1-R. \end{cases} \tag{19}$$

This rate function need not be convex, and whichever of I^N or I^U is smaller dictates whether the ML decoding is the true code-word or a non-transmitted one. The average number of guesses until GRAND identifies a decoding is governed by the beyond-capacity code-book rate,

$$\lim_{n\to\infty} \frac{1}{n} \log E(D^n) = 1 - R.$$

Proof: As $\{1/n \log G(N^n)\}$ and $\{1/n \log U^n\}$ are independent processes, $\{(1/n \log G(N^n), 1/n \log U^n)\}$ satisfies the LDP with rate function $I^N(x) + I^U(y)$. The LDP for $\{1/n \log D^n = 1/n \log \min(G(N^n), U^n)\}$ follows from an application of contraction principle, [3, Th. 4.2.1], with the continuous function $f(x, y) = \min(x, y)$, giving

$$\begin{aligned} I^{\mathrm{D}}(x) &= \inf_{a,b} \left\{ I^N(a) + I^U(b) : f(a,b) = \min(a,b) = x \right\} \\ &= \min\left\{ I^N(x) + \inf_{y\geq x} I^U(y), \inf_{y\geq x} I^N(y) + I^U(x) \right\} \\ &= \min\left\{ I^N(x), \inf_{y\geq x} I^N(y) + I^U(x) \right\}, \end{aligned} \tag{20}$$

where the last line follows from the form of I^U in equation (10).

The simplification of equation (20) into (18) and (19) come about owing to considerations from the following structure. By Theorem 1, the noise guessing rate function starts at the min-entropy rate, $I^N(0) = H_{\min}$. As the min-entropy rate is always less than or equal to the Shannon rate, $H_{\min} \leq H$, $I^N(H) = 0$ and I^N is convex, I^N cannot lie above line from $(0, H_{\min})$ to $(H, 0)$.

If $R < 1 - H$, then $H < 1 - R$ and $I^N(x) \leq I^U(x)$ for all $x \leq H$ from the definition of I^U in equation (10). For $H \leq x \leq 1 - R$, I^N is non-decreasing and so $\min(I^N(x), \inf_{y\geq x} I^N(y) + I^U(x)) = I^N(x)$.

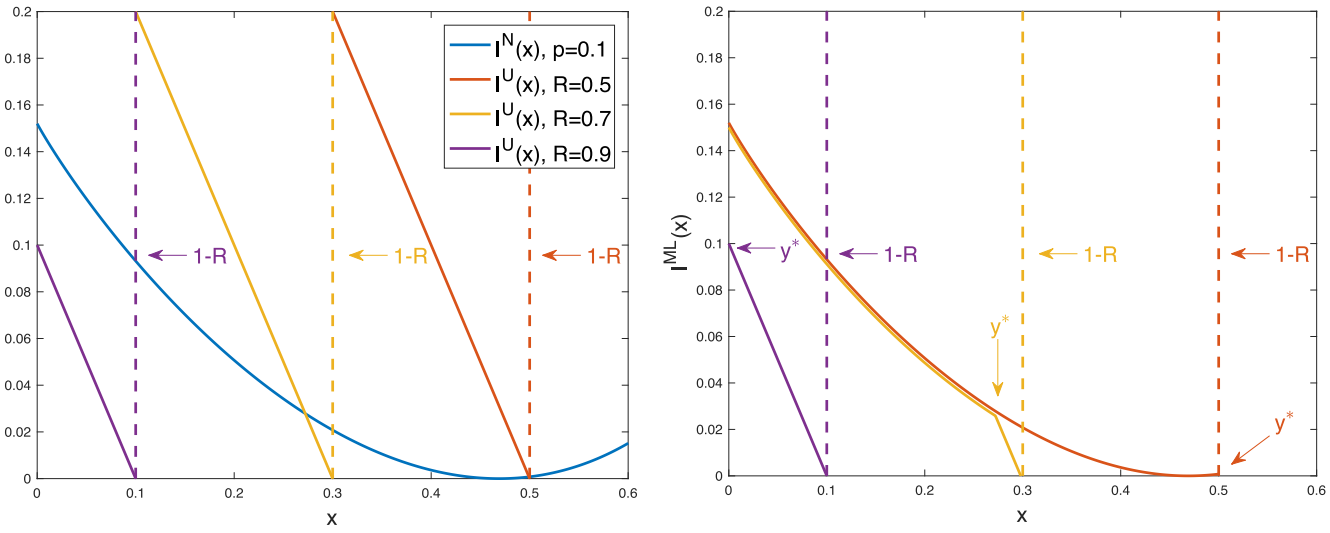

Fig. 4. GRAND complexity. Example: $\mathbb{A} = \{0, 1\}$, BSC channel, noise N^n made of i.i.d. Bernoulli symbols with $P(N^1 = 1) = 0.1$, and channel capacity is approximately 0.53. Code-book consisting of $M_n \approx 2^{nR}$ code-words, uniformly selected in $\mathbb{A}^n$. Left panel: rate function, I^N defined in equation (8), for the number of guesses until the noise is identified $\{1/n \log G(N^n)\}$. Also plotted is the rate function I^U defined in equation (10) for the number of guesses until a non-transmitted element of the code-book is identified, $\{1/n \log U^n\}$. Vertical dashed lines indicate that $I^U(x) = +\infty$ for x to the right of that line. Right panel: as established in Proposition 2, the rate function, I^{D}, that results for the number of queries until an ML decoding is proposed in each of those three cases. Vertical dashed lines indicate that $I^{\mathrm{D}}(x) = +\infty$ for x to the right of that line. If $R < 1 - H$ (red line) so that the code-book rate is within capacity, the zero of I^N occurs before the zero of I^U and the ML decoding mimics the number of guesses until the transmitted word is identified, but with the rate function curtailed at $1 - R$. With $1 - H < R < 1 - H_{\min}$ (yellow line), if the algorithm completes before x^* such that $I^N(x^*) = I^U(x^*)$, whose likelihood is decaying exponentially in n, the true code-word dominates, but ultimately a non-sent code-word is returned. If $R > 1 - H_{\min}$ (purple line), then in this limit, an erroneous code-word is always returned. The super-critical guessing point y^*, which is the supremum over all y satisfying the conditions of Theorem 3, marks the greatest threshold below which, should the ML algorithm declare a decoding has been found, in the large block-length limit, it will be correct, even if the code-book rate is greater than capacity.

If, instead, $R > 1-H$, then $1-R < H$ and $\inf_{y \geq x} I^N(y) = 0$ for all $x \leq 1 - R$, so that $I^{\mathrm{D}}(x) = \min\left\{I^N(x), I^U(x)\right\}$.

To obtain the scaling result for $E(D^n)$ we reverse the transformation from the rate function I^{D} to its Legendre-Fenchel transform, the scaled cumulant generating function of the process $\{n^{-1} \log D^n\}$ via Varadhan's Theorem [3, Th. 4.3.1]. In particular, note that, regardless of whether I^{D} is convex or not,

$$\lim_{n\to\infty} \frac{1}{n} \log E(D^n) = \lim_{n\to\infty} \frac{1}{n} \log E\left(|\mathbb{A}|^{\log D^n}\right) = \sup_{x\in\mathbb{R}} \left\{x - I^{\mathrm{D}}(x)\right\}.$$

If $R < 1-H$, this equals $\min(H_{1/2}, 1-R)$, while if $R > 1-H$ it equals $1 - R$. ∎

One interpretation of the first part of that proposition is that, if the code-book is such that $R < 1 - H$, and so within capacity, identification of the correct code-word occurs because it is likely that all elements in the typical set of the noise will be queried before a non-transmitted element of the code-book is identified. Owing to the long tail of guesswork, in the absence of the other elements, of the code-book stopping the guessing algorithm, the average number of guesses that would be made would grow with rate $H_{1/2}$ [15]. If, however, one minus the normalized code-book rate R is less than that, the long tail of the scheme is clipped. While this clipping is not enough to make an error likely, it is enough to reduce the average number of queries that will be made before an element of the code-book is identified.

Example. *An example of the range of behaviors described in Proposition 2 for a BSC can be found in Fig. 4. The non-convex rate function corresponds to a code-book rate beyond capacity, $R > 1 - H$.*

If the code-book rate is beyond capacity, $R > 1 - H$, then implicit in the results of Proposition 2 is that there are circumstances where, conditioned on the unlikely event that the algorithm terminates after a relatively small, but exponentially growing, number of guesses, the decoded code-word GRAND identifies is certain to be the transmitted code-word in the large block length limit. While this property can appear under more nuanced circumstances, we provide one condition where the resulting characterization is simple. Namely if the code-book rate is between channel capacity and one minus the min-entropy rate of the noise, $1 - H < R < 1 - H_{\min}$, then one can determine an exponent below which, in the limit as the block length becomes large, if the ML algorithm terminates after a number of guesses below the threshold governed by that exponent, the decoded code-word will correctly correspond to the transmitted code-word.

Theorem 3. *Under Assumptions 1 and 2, if $0 < y < 1 - R$ is such that $I^N(y) < I^U(y)$, then the probability of a correct decoding given fewer than $|\mathbb{A}|^{ny}$ queries are made before the algorithm terminates converges to 1,*

$$\lim_{n\to\infty} P\left(G(N^n) < U^n \middle| \frac{1}{n} \log D^n \leq y\right) = 1.$$

Such a y necessarily exists if the code-book rate is less than one minus the min-entropy rate of the noise, $R < 1 - H_{\min}$.

Proof: To see that such a y exists if $R < 1-H_{\min}$, observe that as $R < 1 - H_{\min}$ we have that the noise guessing rate function starts strictly below the non-transmitted guessing rate function, $I^N(0) = H_{\min} < 1 - R = I^U(0)$. As both I^N and I^U are continuous, the existence of such a y is guaranteed.

Defining the continuous function $f : [0,1]^2 \to [0,1]^3$ by $f(x,y) = (x, y, \min(x,y))$, then by the contraction principle,

$$\left\{\left(\frac{1}{n}\log G(N^n), \frac{1}{n}\log U^n, \frac{1}{n}\log D^n\right)\right\}$$

satisfies the LDP with rate function

$$I^{N,U,D}(x,y,z) = \begin{cases} I^N(x) + I^U(y) & \text{if } z = \min(x,y) \\ +\infty & \text{otherwise .} \end{cases}$$

We apply the [68, Th. 3.1] to establish the concentration of measure conditioned on the rare event that the algorithm terminates within $|\mathbb{A}|^{ny}$ guesses. By that theorem, we have that for any open neighborhood B of $(\min(y,H), 1-R, \min(y,H))$,

$$\lim_{n\to\infty} P\left(\left(\frac{\log G(N^n)}{n}, \frac{\log U^n}{n}, \frac{\log D^n}{n}\right) \in B \,\middle|\, \frac{\log D^n}{n} \le y\right) = 1,$$

from which the result follows. ∎

If the code-book rate is less than capacity, Theorem 3 recovers what we already knew from Proposition 1: that we have concentration of measure onto correct decodings. Even if the code-book rate is beyond capacity, however, it establishes that, conditioned on the algorithm terminating early, there are circumstances where we shall have concentration onto correct decodings. Examples to this effect are presented in the right hand panel of Fig. 4, where the supremum over all y satisfying the condition of Theorem 3, y^*, which we call the super-critical guessing threshold, is marked. For code-book rates that are greater than capacity, i.e. the left two lines, $y^* < H$ and the ML decoding is only likely to be correct if the GRAND algorithm terminates in a number of queries in the guesswork order that is below approximately $|\mathbb{A}|^{ny^*}$.

B. Approximate ML Decoding With GRANDAB

While Proposition 2 identifies the computational complexity of GRAND and so is directly related to the decoding algorithm, Proposition 1 provides a version of the Channel Coding Theorem for ML decoding in general. That is, it relates to the likelihood that an ML decoding is in error, irrespective of the algorithm used to identify the ML decoding. Its proof via noise guessing, however, suggests an approximate ML decoding scheme, GRANDAB, with constrained complexity.

If the code-book rate is within capacity, $R < 1 - H$, the likelihood of erroneous decoding is strictly decaying in n. Essentially this occurs as the likelihood of identifying a transmitted noise sequence is dominated by queries to up to, and including, the Shannon Typical Set, a fact made clear by $I^N(H) = 0$. The expected guessing location to the first non-transmitted element encountered is governed by one minus the code-book rate, $I^U(1-R) = 0$. Thus when $R < 1 - H$, $H < 1 - R$ and guessing the true input dominates over identifying a non-transmitted code-word.

That guessing the noise has a long tail beyond H is a consequence of large growth in the number of sequences to be queried when compared to the rate of acquisition of probability on querying them, leading to the undesirable $H_{1/2}$ growth rate for unconstrained noise guessing. For dense code-books, this guessing tail is clipped with an error at $1 - R$, but - despite that error - capacity is achieved so long as the code is within capacity $R < 1 - H$. Further contemplation of this fact suggests the following algorithm: perform the GRAND, but abandon guessing after $|\mathbb{A}|^{n(H+\delta)}$ queries, for some $\delta > 0$, declaring an error. This algorithm does not implement ML decoding, but it is still capacity achieving.

Proposition 3. *(GRANDAB Coding Theorem and Guessing Complexity). Under the assumptions of Theorems 1 and 2. If the code-book rate is less than the capacity, $R < 1 - H$, then the GRANDAB error rate is*

$$\lim_{n\to\infty} \frac{1}{n}\log P\left(\left\{U^n \le G(N^n)\right\} \cup \left\{\frac{1}{n}\log G(N^n) \ge H + \delta\right\}\right) = -\min\left\{\inf_{a\in[H,1-R]}\{I^U(a) + I^N(a)\}, I^N(H+\delta)\right\} < 0,$$

so that probability that the ML decoding is not the transmitted code-word decays exponentially in the block length n. If, in addition, x^* defined in equation (12) *exists then this simplifies to what we call the GRANDAB error rate*

$$\epsilon^{AB}(R) = \min\left(\epsilon(R), I^N(H+\delta)\right) \tag{21}$$

where $\epsilon(R)$ is the ML decoding error rate in equation (13)*. The expected number of guesses until GRANDAB terminates, $\{D^n_{AB}\}$, satisfies*

$$\lim_{n\to\infty} \frac{1}{n}\log E(D^n_{AB}) = \min\left(H_{1/2}, 1-R, H+\delta\right).$$

For rates above capacity, $R > 1 - H$, the success probability is identical to that for ML decoding, given in equation (14)*.*

Proof: By the principle of the largest term, [3, Lemma 1.2.15] or [69],

$$\limsup_{n\to\infty} \frac{1}{n}\log P\left(\left\{U^n \le G(N^n)\right\} \cup \left\{\frac{1}{n}\log G(N^n) \ge H+\delta\right\}\right) = \max\left(\limsup_{n\to\infty}\frac{1}{n}\log P\left(U^n \le G(N^n)\right), \limsup_{n\to\infty}\frac{1}{n}\log P\left(\frac{1}{n}\log G(N^n) \ge H+\delta\right)\right),$$

with a similar equation holding for lim inf. The behavior of the first term is identified in Proposition 1. The behavior of the second term is established directly from the LDP in Theorem 1 on noting that

$$\inf_{x\ge H+\delta} I^N(x) = I^N(H+\delta),$$

as I^N is strictly increasing beyond H. Coupled with the continuity of I^N, we obtain equation (21). The expected number of guesses until the algorithm completes is determined in an identical manner to that in Proposition 2. ∎

The interpretation of this result is straight-forward: GRANDAB results in an error if either the ML decoding

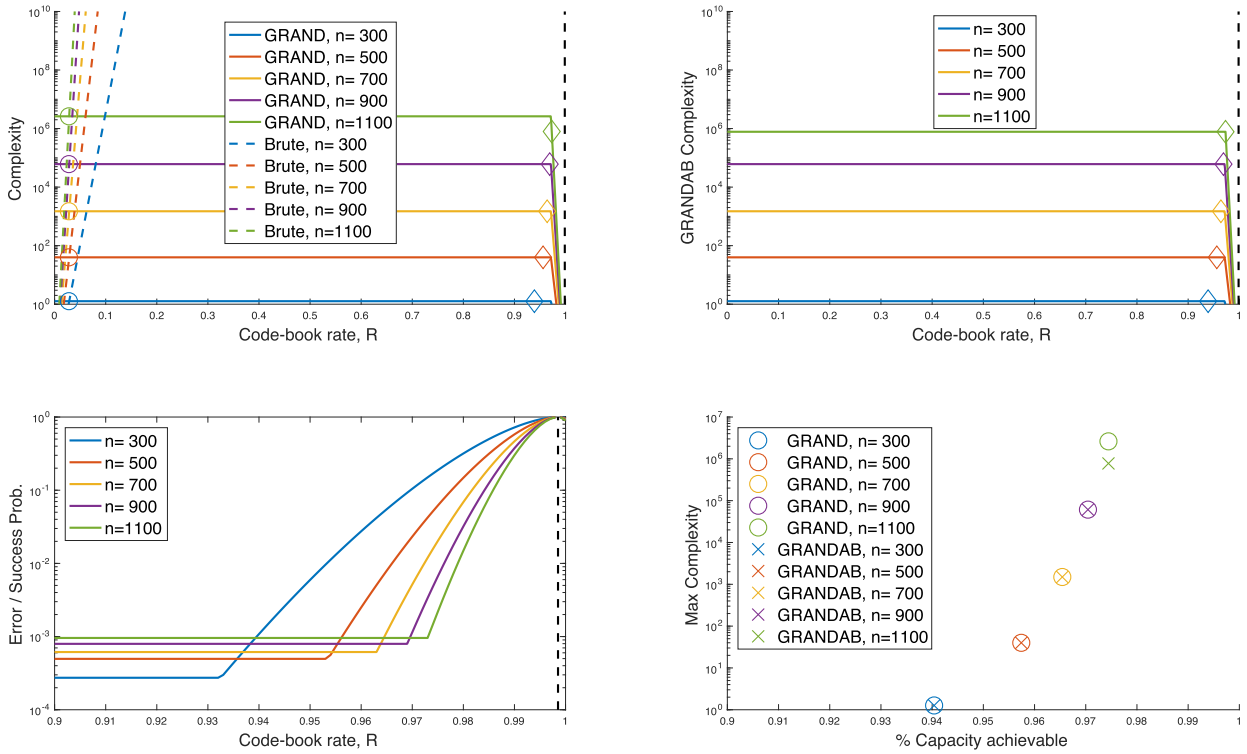

Fig. 5. BSC GRAND and GRANDAB decoding. Bit flip probability $p = 10^{-4}$, code-book rate R and block length n. Dashed vertical lines in three of the panels indicate channel capacity. Top left panel: complexity of ML decoding by noise guessing (solid lines) or by brute force (dashed lines) as a function of code-book rate. Circles indicate the rate beyond which computing within the code-book has higher complexity than noise guessing. Diamonds indicate the rate below which block error probability is less than 10^{-3}. Top right panel: complexity of GRANDAB as a function of code-book rate, where the free parameter δ in GRANDAB is selected as described in Section IV. The inflection in complexity in these top two panels occurs at the cut-off rate. Bottom left panel: with a zoomed in x-scale, to the left of capacity the curves show approximate error probability of GRANDAB for a range of n. To the right of capacity the curves show approximate success probability of both GRAND and GRANDAB. Bottom right panel: for each block length and both GRAND and GRANDAB, the maximum achievable rate, as a percentage of capacity, while keeping the block error probability below 10^{-3} is plotted against the highest complexity of the code, which occurs for low-rate code-books.

is erroneous, as governed by Proposition 1, or if the algorithm abandons guessing before an element of the code-book is identified. Whichever of these two events is more likely dominates the error rate. So long as the algorithm does not abandon until after querying all elements in the typical set of the noise, it is capacity achieving.

The earlier Theorem 3 also suggests an abandonment rule when code-books are at rate beyond capacity. One could curtail querying and declare an error after approximately $|\mathbb{A}|^{ny^*}$ guesses, where y^* is maximum over all y satisfying the conditions of Theorem 3. Before that point, it is likely that the decoding is correct, while afterwards it is likely to be incorrect.

IV. Examples

As all of the results in this paper hold for channels with memory, to illustrate the complexity, error and success probabilities of GRAND and GRANDAB decoding we treat binary $\mathbb{A} = \{0, 1\}$ noise sequences $\{N^n\}$ whose elements are chosen via a Markov chain with transition matrix

$$\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix},$$

and assume that $1 > a, b > 0$. The initial distribution of the Markov chain can go unspecified as it plays no role in the asymptotic results. This model includes the BSC by setting $p = a = 1 - b$, but, in general, the second eigenvalue is $1-a-b$, which characterizes the burstiness, memory or mixing of the Markov chain.

The Rényi entropy rate of this noise source can be evaluated [16] for $\alpha \neq 1$ to be

$$H_\alpha = \frac{1}{1-\alpha} \log \left((1-a)^\alpha + (1-b)^\alpha + \sqrt{((1-a)^\alpha - (1-b)^\alpha)^2 + 4(ab)^\alpha} \right) - \frac{1}{1-\alpha}.$$

While with $h(a) = -a \log(a) - (1-a) \log(1-a)$ being the binary Shannon entropy, $H_1 = H = h(a)b/(a+b) + h(b)a/(a+b)$ is the Shannon entropy rate of the Markovian source. Thus using equation (6) we have an explicit expression for the resulting scaled cumulant generating function, Λ^N, of the logarithm of the noise. While the rate function I^N defined in equation (8) cannot be calculated in closed form, it is readily evaluated numerically, only requiring the solution of a one-dimensional concave optimization.

While prefactors are not captured in that asymptotic analysis in Propositions 1, 2 and 3, they allow the following approximations. For GRAND and GRANDAB decoding, our measure

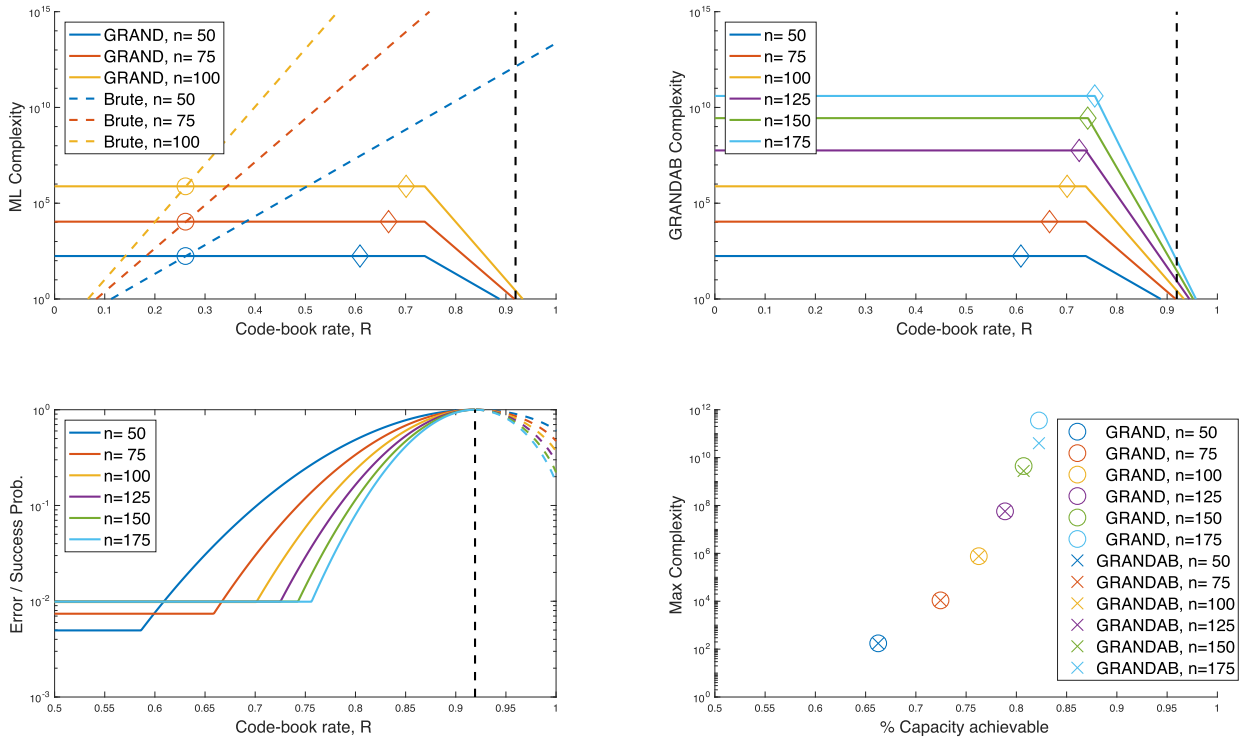

Fig. 6. BSC GRAND and GRANDAB. Same display as for Fig. 5, but with bit flip probability $p = 10^{-2}$ and a block-error probability floor of 10^{-2}.

of complexity is the average number of guesses per bit:

$$\text{GRAND ave. no. guesses / bit} \approx \frac{2^{n \min(1-R, H_{1/2})}}{n}$$

$$\text{GRANDAB ave. no. guesses / bit} \approx \frac{2^{n \min(1-R, H_{1/2}, I^N(H+\delta))}}{n}.$$

For comparison, we define the complexity of the straight computation of the ML decoding in (4) to be the number of conditional probabilities that must be computed per bit before rank ordering and determining the most likely code-book element:

$$\text{No. conditional prob. computations / bit} = \frac{2^{nR}}{n}.$$

Thus we are equating the work performed in one noise guess with one computation of a conditional probability. As this direct scheme results in the ML decoding as by noise guessing, it shares the same error and success probabilities as GRANDAB.

For error and success probabilities we employ: GRAND probability of error $\approx 2^{-n\epsilon(R)}$ for $R < 1 - H$; GRANDAB probability of error $\approx 2^{-n\epsilon^{\text{AB}}(R)}$ for $R < 1 - H$; GRAND & GRANDAB probability of success $\approx 2^{-ns(R)}$ $R > 1 - H$; where ϵ, ϵ^{AB}, and s are given in equations (13), (21) and (14).

We use the following rule to select the parameter δ that determines how far beyond the Shannon typical set queries are made before abandonment in GRANDAB. With the stationary probability of noise per bit being p, for a given block-length n we identify δ such that the probability of abandonment is no more than p_{abandon} times the expected uncoded block error probability; i.e we solve the following equation numerically for $\delta(n)$:

$$2^{-nI^N(H+\delta(n))} = p_{\text{abandon}} \min(pn, 1).$$

Selecting this δ sets a floor for the block-error probability generated by abandoned guessing that is a fraction of the uncoded block-error probability.

We set $p_{\text{abandon}} = 10^{-3}$ if the average bit error rate in the channel is 10^{-4} and $p_{\text{abandon}} = 10^{-2}$ if it is 10^{-2} indicating we are willing to tolerate block-error probabilities that are of order at least 100 or 1000 times less likely than an uncoded block error.

For complexity, as the number of computations per bit per second is normally several orders of magnitude greater than the number of bits received over the channel per second, we will consider a complexity feasible if it is in the range of $10^3 - 10^4$ guesses per bit. For both GRAND and GRANDAB, this is likely to be a conservative constraint as the guessing is readily parallelizable.

A. Binary Symmetric Channel (BSC)

For the BSC with bit error probability $p = 10^{-4}$, a GRANDAB decoding abandonment probability of $p_{\text{abandon}} = 10^{-3}$, and a range of block lengths, n, the approximate complexity and error performance of GRAND and GRANDAB is shown in Fig. 5.

The top left panel shows the complexity (average number of guesses per received bit) for GRAND (solid lines) and by brute force (dashed lines) for a range of block lengths, n, with the vertical dashed line indicating capacity, $1 - H$. The computational complexity of the brute force approach, computing conditional probabilities for all elements of the code-book rapidly grows with rate. The complexity of guessing

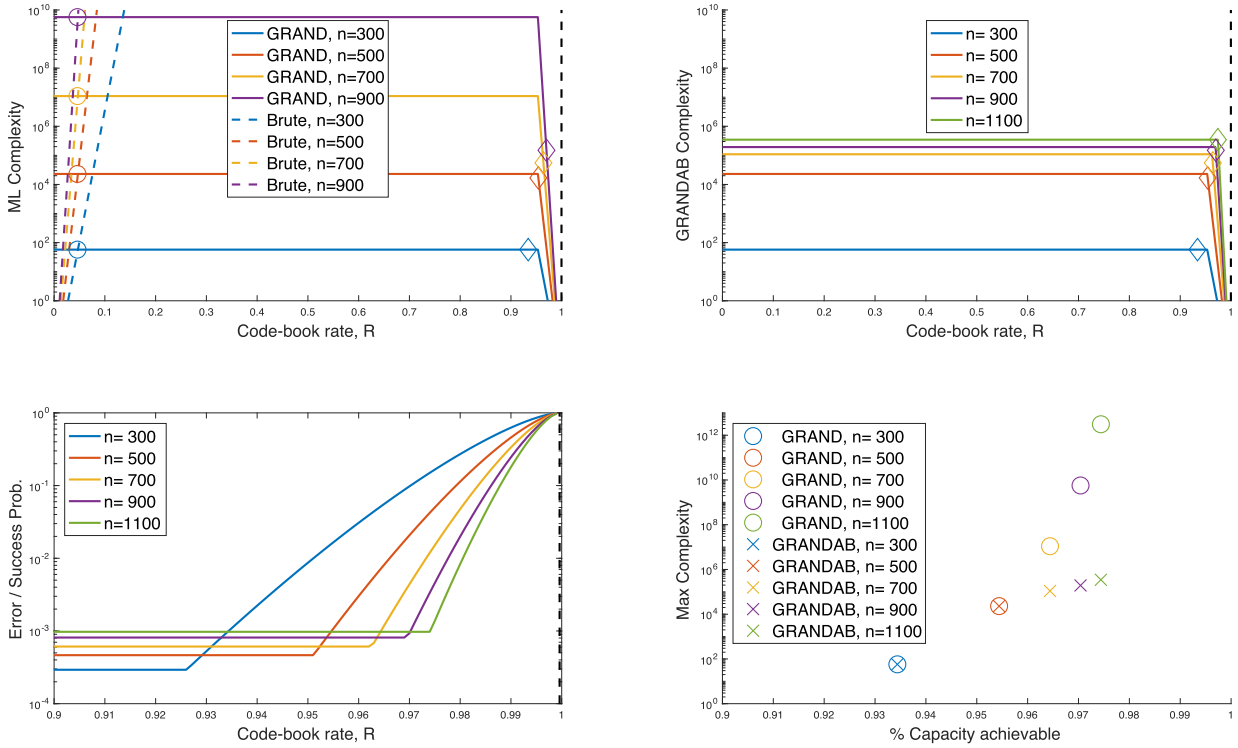

Fig. 7. GRAND and GRANDAB decoding with binary Markovian noise. Average bit flip probability $p = 10^{-4}$, making it comparable to the BSC plots in Fig. 5, but for a Markovian channel with $a = p/5 = 2 \times 10^{-5}$ and $b = (1-p)/pa \approx 0.2$, so making an extremely bursty noise channel. Four displayed panels are analogous to those described in the caption of Fig. 5.

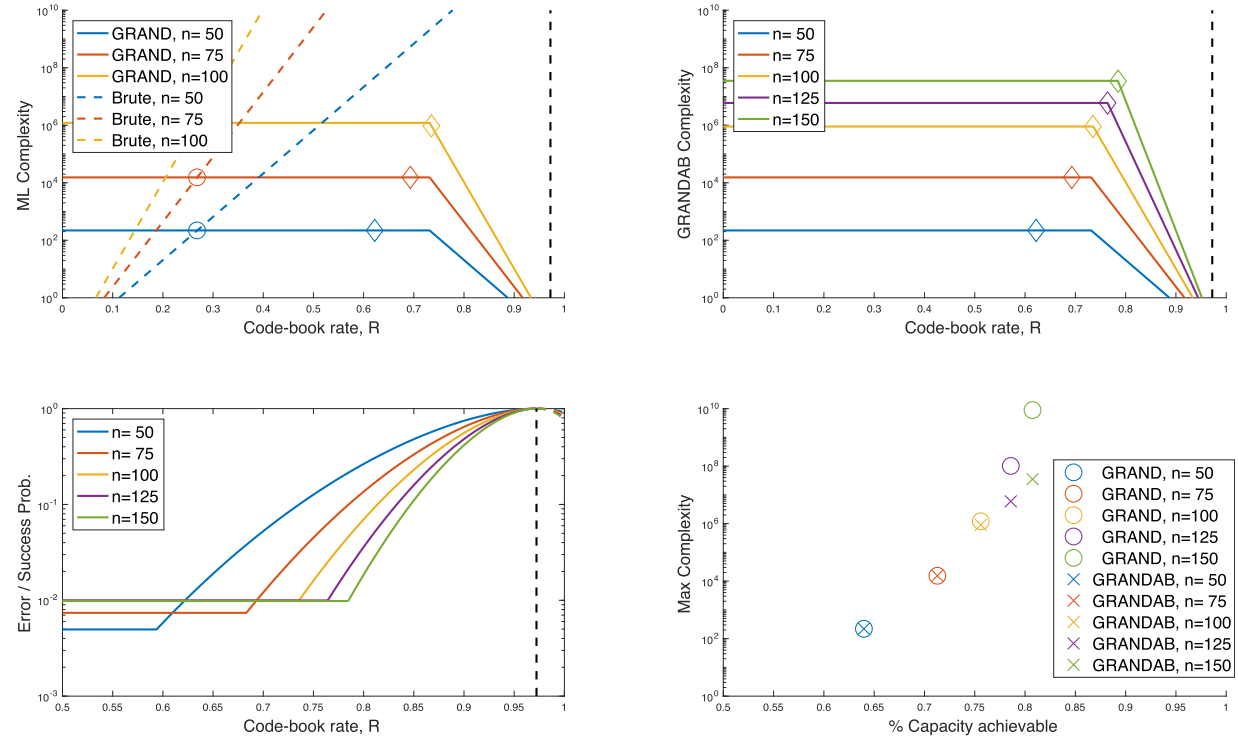

Fig. 8. GRAND and GRANDAB with Markovian noise. Same display as for Fig. 7, but with average bit flip probability $p = 10^{-2}$.

the noise only decreases as rates increase, with the circles indicating the threshold above which the complexity of guessing within the code-book is less than that of brute force determination. The diamond marks the code-book rate after which the block error probability for GRAND is $p_{\text{block}} = 10^{-3}$ and so sets an upper-threshold on the code-book rate. The top right panel shows the equivalent complexity plot for GRANDAB decoding. The effect of abandonment is to reduce the maximum complexity for the longest block-length, with no impact on smaller block-lengths in this instance.

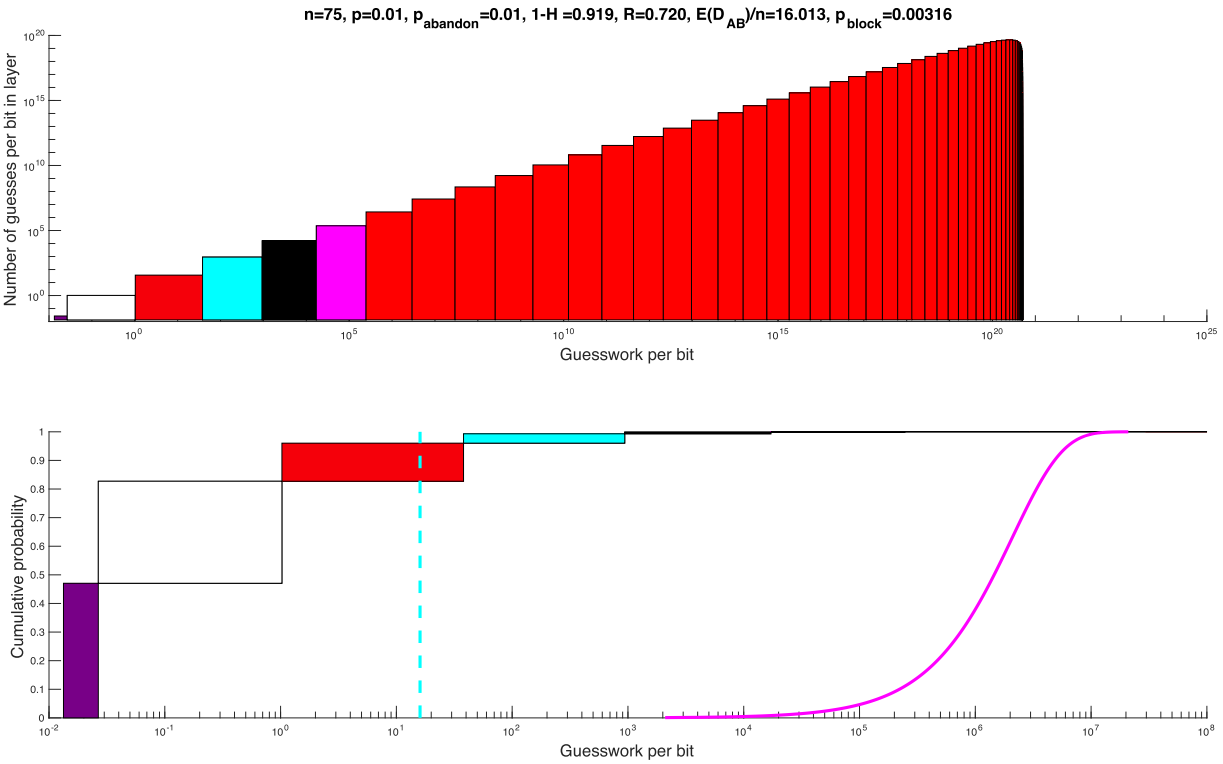

Fig. 9. Example: $\mathbb{A} = \{0, 1\}$, BSC channel, bit flip probability $p = 10^{-2}$, block length $n = 75$, capacity $1 - H = 0.919$ and a code-book rate of $R = 0.72$. Upper panel: The x-axis is the total number of queries per-bit on a log-scale. The y-axis is the number of queries per-bit that are made to sequences of increasing Hamming distance, also on a log-scale. Each rectangle demarcates a distinct Hamming distance. The color coding indicates the probability that is accumulated by guessing through a layer of given Hamming distance and runs from blue, 0, to red, 1. The white layer is where 2^{nH} guesses have been made, at the core of the Shannon Typical Set. Accumulation of probability around the white layer is asymmetric. Prior to it, probability is quickly obtained, but the decreasing probability per sequence coupled with the increasing number of sequences with the same probability results in $2^{nH_{1/2}}$, asymptotically the average guesswork, being in the black layer. The cyan layer indicates the guessing layer by the end of which there is a 99% chance of identifying true noise. Abandoning guessing here if no code-word had been identified would result in 205 fewer guesses per bit than would, on average, be necessary to identify the true noise sequence. With a code-book rate of 0.72, using the approximation in (9), the magenta layer is where a non-transmitted code-word would, on average, be identified. Lower Panel: Cumulative probability of guesswork with the same color coding as the upper panel, but with a truncated y-axis to show more detail. The dashed vertical cyan line is located at the average number of guesses per-bit per GRANDAB decoding, $E(D_{\text{AB}})/n \approx 16$. The magenta line is the cumulative distribution of the number of guesses per bit until a non-transmitted code-word is identified, using the approximation to U^n found in equation (9). Note that, with a log x-scale, it is tightly centered around its mean, resulting in a block error probability of $p_{\text{block}} = 3.15 \times 10^{-3}$.

The bottom left panel shows the approximate block-error and block-success probabilities below and above capacity, respectively, for GRANDAB as a function of code-book rate. The ML curves would be identical at higher rates, but would drop further at lower code-rates as the abandonment of guessing of GRANDAB is what places a floor on the block-error rate.

For both GRAND and GRANDAB, the final panel, bottom right, shows the maximum complexity for a given block length, n, versus the % of capacity achievable with a code-book rate that provides a block error probability below $p_{\text{block}} = 10^{-3}$. With the rule of thumb that $10^3 - 10^4$ guesses per bit is acceptable, then choosing $n = 700$ could realize up to 96.5% of capacity. Note that this occurs for a block length that is substantially smaller than the reciprocal of the bit error rate, $1/p = 10,000$.

The inflection in complexity for the top two panels occurs at the cut-off rate. This illustrates an intriguing property of GRAND and GRANDAB. While for sequential decoding of tree codes, decoding complexity increases steeply when the rate exceeds the cut-off rate, for decoding by guessing noise, complexity decreases past the cut-off rate.

Analogous information is displayed for the BSC with bit error probability $p = 10^{-2}$ in Fig. 6, but with $p_{\text{abandon}} = 10^{-2}$. Again, the computational complexity of the brute force approach makes it infeasible even for modest rates. For these higher bit error probabilities, the effect of GRANDAB's truncation is felt at smaller block sizes. This might be expected, given the Shannon entropy of the noise has increased. As the likelihood of noise is increased, block-lengths must be reduced to keep guesswork down to the 10^3–10^4 guesses per-bit range. For $p = 10^{-2}$, complexity considerations reduce n to 75, for which rates providing up to 72.4% of capacity are achievable with a block error probability no more than $p_{\text{block}} = 10^{-2}$.

B. Bursty Markovian Noise

A core feature of the proposed schemes is that they can be applied in channels with correlated noise without the need for interleaving and other methods that attempt to mask the impact of memory. The equivalent of Fig. 5 is presented in Fig. 7 where the long run average probability of bit-error is set to be the same, $p = 10^{-4}$, in both, but here $a = 10^{-4}/5$ and $b = 1/5$. These have been selected to give a highly bursty source where the likelihood of a bit flip is small, but the

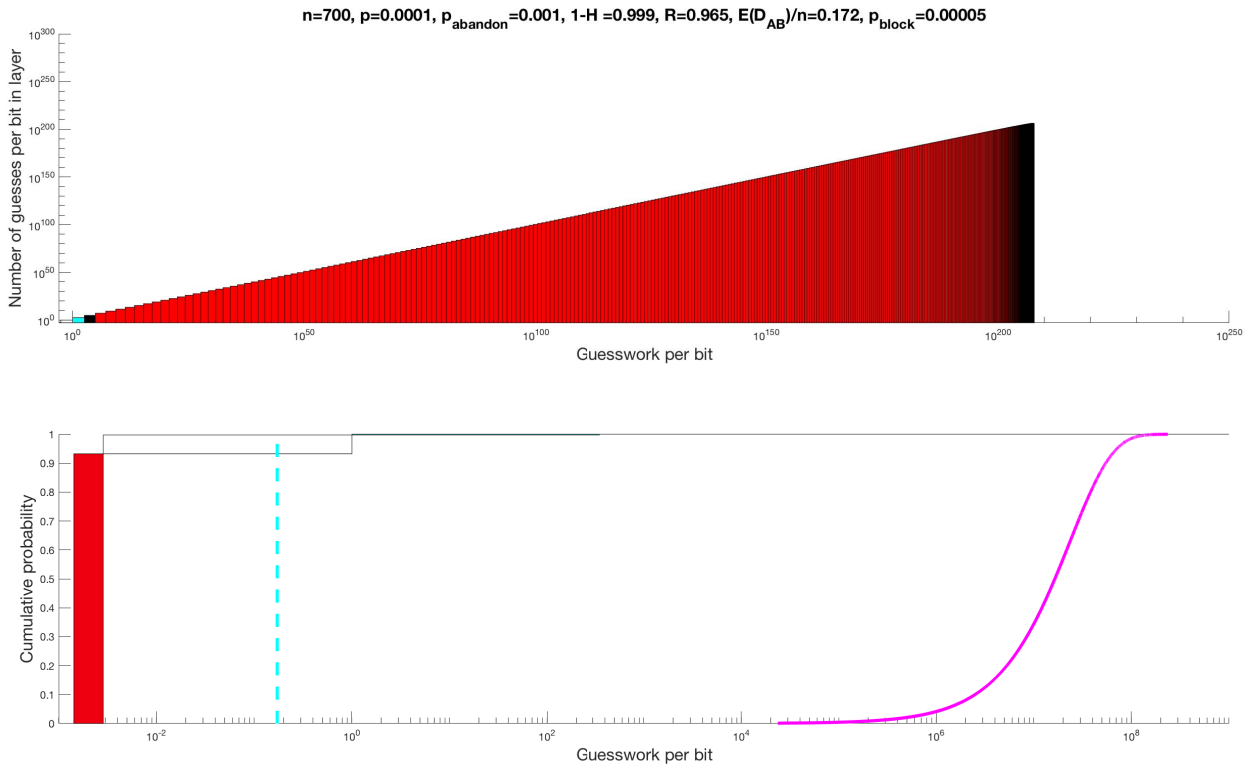

Fig. 10. Example: $\mathbb{A} = \{0, 1\}$, BSC channel, bit flip probability $p = 10^{-4}$, block length $n = 700$, capacity $1 - H = 0.999$ and a code-book rate of $R = 0.965$. Upper panel: Same layout as in Fig. 9. Abandoning guessing after 99.9% probability is accumulated if no code-word had been identified would result in 115 fewer guesses per bit than would, on average, be necessary to identify the true noise sequence. Lower Panel: As in Fig. 9. Average number of guesses per-bit per GRANDAB decoding $E(D_{\text{AB}})/n \approx 0.172$. Block error probability of $p_{\text{block}} = 4.69 \times 10^{-5}$.

likelihood of an additional bit flip given one has occurred is 3 orders of magnitude higher. The block-lengths displayed for the Markovian channels are the same as for the corresponding BSC example and again $p_{\text{abandon}} = 10^{-3}$, to enable ready comparison.

For this parameterization, the complexity of GRAND is much higher for this Markovian noise than the BSC equivalent. Consequently, GRANDAB plays a more significant role in reducing that complexity for large block lengths by abandonment. Based on the criteria set for the BSC, for reasons of complexity $n = 500$ would be selected. While this is shorter than the block length for the equivalent BSC, it is still the case that 95.4% of capacity is achievable with a block error rate of less than $p_{\text{block}} = 10^{-3}$.

Fig. 8 can be compared with the BSC in Fig. 6, having $p = 10^{-2}$ obtained by $a = 10^{-2}/5$ and $b = a(1 - p)/p \approx 0.198$. For this noisy channel, again GRANDAB provides a reduction in algorithmic complexity at a cost of introducing an error floor. Limit complexity at the receiver, one would select $n = 75$. With a threshold of a block-error rate set at 10^{-2}, 71.2% of capacity is available.

Note that, in all examples presented here, the best block lengths are no larger than the reciprocal of the corresponding bit error rate, $1/p$. This behavior may be unexpected if we consider error exponents for Markov channels based on interleaving of the order of the mixing time of the Markov noise model [70], yet it is a desirable feature of the scheme, which we have consistently observed.

C. Finer Approximations for the BSC

For uniform-at-random code-books, Proposition 1 provides error exponents for general noise processes. In the case of the memoryless channel, however, a more exact computation of the block error probability is possible. This is achieved by availing of the precision of the finer approximation to the distribution of the number of guesses until a non-transmitted code-word is identified, U^n, given in equation (9).

The error probability is one minus the success probability,

$$P(U^n \leq G(N^n)) = 1 - P(G(N^n) < U^n),$$

and we shall provide a more exact computation of the latter. Restricting to a BSC, there are n choose 0 noise strings with no errors, n choose 1 strings with one error, and so forth. Thus we define $l_{-1} = 0$ and

$$l_k = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k}$$

for each $k \in \{0, \ldots, n\}$. Consequently in guesswork order we have

$$P(G(N^n) = m) = p^k (1 - p)^{n-k}$$

for every $m \in \{l_{k-1} + 1, \ldots, l_k\}$. Thus

$$P(G(N^n) < U^n) = \sum_{k=0}^{n} p^k (1-p)^{n-k} \sum_{m=l_{k-1}+1}^{l_k} P(U^n > m).$$

Approximating the distribution of U^n by

$$P(U^n > m) \approx \exp(-m2^{-n(1-R)}),$$

as suggested by equation (9), and computing the resulting geometric sum gives

$$P(G(N^n) < U^n) \approx \sum_{k=0}^{n} p^k (1-p)^{n-k} \left(\frac{e^{-(l_{k-1}+1)2^{-n(1-R)}} - e^{-(l_k+1)2^{-n(1-R)}}}{1 - e^{2^{-n(1-R)}}} \right).$$

Thus, for a BSC we can compute a finer approximation to the block error probability, p_{block}, by a sum of only $n+1$ terms.

Fig. 9 reconsiders the scenario treated via the large deviations analysis in Fig. 6, but with this finer approximation for the block error probability. The n and R used correspond to those deduced from the asymptotic analysis as maximizing rate subject to constraints on block error probability while maintaining a certain degree of complexity. The true block error probability is 3×10^{-3}, when the target in the asymptotic regime was 10^{-2} indicating good accuracy.

In all cases we have examined beyond those shown here, the asymptotic results compare well with the more precise computations which, if anything, suggest that higher rates can be obtained while still meeting block error targets.

V. Discussion and Conclusions

We have introduced and analyzed two decoding algorithms based on guessing that are suitable for a broad class of noise processes. Subtracting noise from a received signal in order from most likely to least likely, the first instance that is in the code-book corresponds to the ML decoding. Both GRAND, which identifies an ML decoding by noise guessing, and GRANDAB, an approximate ML decoding by noise guessing algorithm in which the receiver quits its attempts to identify an element of the code-book after a given number of unsuccessful queries that is determined by the Shannon entropy of the noise, are capacity achieving when used with uniform-at-random code-books. To establish capacity results, we have assumed that the source is uniform. Depending on channel conditions, GRANDAB has the potential benefit over ML decoding of decreased complexity, even for DMCs. Analytically leveraging this noise-focused view, we provide explicit error and success exponents for code-book rates that are within and beyond capacity, respectively, providing a version of the Channel Coding Theorem.

While DMCs form the classic model in information theory, real communication channels are not memoryless, e.g. [71], and commonly are made artificially so by interleaving for many existing decoding schemes to function well, leading to additional delays in encoding and decoding. In contrast, all of the results presented in the present paper for GRAND and GRANDAB hold directly for noise processes with more involved structures, and no interleaving is required for their use. To illustrate that, we have presented analytic examples based on bursty Markovian noise.

The noise guessing approach underlying GRAND and GRANDAB has other desirable features. For example, both schemes are universally applicable in the sense that their execution only depends on the structure of the noise rather than how the code-book was constructed. Moreover, guesswork orders are known to be robust to mismatch [63].

For both GRAND and GRANDAB, we provide asymptotic results on the number of queries that the receiver must make per received code-word for uniform-at-random code-books. Notably, the approach becomes less complex as the code-book rate increases.

While testing a string's membership of a uniform-at-random code-book can be achieved efficiently with the code-book stored in a $\mathbb{A}$-ary tree, the code-book description requires substantial memory, limiting utility for large block-lengths. Any use of a random code-book also requires techniques for encoding, and for converting a code-word to an information word, but both of these can be performed with linear complexity. To encode, potential inputs can be stored in a lexicographically ordered $\mathbb{A}-$ary tree of depth nR with a final leaf that contains a string of length n, the code-word to be transmitted. Thus finding an encoding entails a tree search, i.e. nR operations. When mapping a code-word to an information word, the code-book can be stored as a lexicographically ordered $\mathbb{A}-$ary tree of depth n with a final leaf that contains a string of length nR, the corresponding information-word. Thus, identifying an information word also requires a tree search, i.e. n operations.

An alternative instantiation of the schemata would be realized in combination with linearly constructed code-books such as Hamming, LDPC, or random linear code-books. While binary linear code-books can be capacity achieving for the BSC [44], random linear code-books have recently been shown to be capacity-achieving [39] for DMCs. To describe a linear code-book, one need solely record its generator matrix and so storage is small. Using the parity check matrix associated with the generator, testing a string for membership of the code-book is efficient as it only requires the computation of the syndrome of the received string less guessed noise. Using ML decoding by noise guessing with linear code-books effectively results in replacing the usual coset leader of each syndrome, the noise string in the coset with minimum Hamming weight, with the most likely noise string in the coset. A thorough investigation of that possibility, along with small block size properties, integration into outer coding schemes, and so forth, is the topic of ongoing work. The current work treats only a hard detection model where only discrete data is presented to the decoder. Extending the principles of these noise guessing techniques to a continuous case where soft detection information is available imputes quantization issues that merit their own investigation, and is the subject of further ongoing work.

References

[1] M. M. Christiansen and K. R. Duffy, "Guesswork, large deviations, and Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 59, no. 2, pp. 796–802, Feb. 2013.

[2] M. M. Christiansen, K. R. Duffy, F. du Pin Calmon, and M. Médard, "Multi-user guesswork and brute force security," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6876–6886, Dec. 2015.

[3] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. New York, NY, USA: Springer-Verlag, 1998.

[4] V. Anantharam, "A large deviations approach to error exponents in source coding and hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 36, no. 4, pp. 938–943, Jul. 1990.

[5] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York, NY, USA: Academic, 1981.

[6] L. Davisson, G. Longo, and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic Markov sources," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 4, pp. 431–438, Jul. 1981.

[7] A. Dembo and L. Kontoyiannis, "Source coding, large deviations, and approximate pattern matching," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1590–1615, Jun. 2002.

[8] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 1, pp. 3–18, Jan. 1965.

[9] A. Montanari and G. D. Forney. (2001). *On Exponential Error Bounds for Random Codes on the DMC*. [Online]. Available: https://web.stanford.edu/~montanar/RESEARCH/FILEPAP/dmc.ps

[10] R. Yaguchi, V. Y. F. Tan, S. Watanabe, and M. Hayashi. (2017). *Large and Moderate Deviations for Joint Source-Channel Coding of Systems With Markovian Memory*. [Online]. Available: https://www.ece.nus.edu.sg/stfpage/vtan/YTWH17.pdf

[11] Y. Zhong, F. Alajaji, and L. L. Campbell, "On the joint source-channel coding error exponent for discrete memoryless systems," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1450–1468, Apr. 2006.

[12] I. Csiszár, "Joint source-channel error exponent," *Problems Control Inf. Theory*, vol. 9, no. 5, pp. 315–328, 1980.

[13] I. Csiszár, "On the error exponent of source-channel transmission with a distortion threshold," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 6, pp. 823–828, Nov. 1982.

[14] J. L. Massey, "Guessing and entropy," in *Proc. IEEE Int. Symp. Inf Theory*, Jun./Jul. 1994, p. 204.

[15] E. Arikan, "An inequality on guessing and its application to sequential decoding," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 99–105, Jan. 1996.

[16] D. Malone and W. G. Sullivan, "Guesswork and entropy," *IEEE Trans. Inf. Theory*, vol. 50, no. 3, pp. 525–526, Mar. 2004.

[17] C. E. Pfister and W. G. Sullivan, "Rényi entropy, guesswork moments, and large deviations," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2794–2800, Nov. 2004.

[18] E. Arikan, "Large deviations of probability rank," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2000, p. 27.

[19] E. Arikan and N. Merhav, "Guessing subject to distortion," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1041–1056, May 1998.

[20] M. K. Hanawal and R. Sundaresan, "Guessing and compression subject to distortion," Division Elect. Sci., Indian Inst. Sci., Bengaluru, India, Tech. Rep. TR-PME-2010-12, 2010.

[21] P. Elias, "Error-free Coding," *Trans. IRE Prof. Group Inf. Theory*, vol. 4, no. 4, pp. 29–37, Sep. 1954.

[22] J. M. Wozencraft, "Sequential decoding for reliable communication," Ph.D. dissertation, Dept. Elect. Eng., MIT, Cambridge, MA, USA, 1957.

[23] J. M. Wozengraft and B. Reiffen, *Sequential Decoding*. Cambridge, MA, USA: MIT Press, 1961.

[24] R. Fano, "A heuristic discussion of probabilistic decoding," *IEEE Trans. Inf. Theory*, vol. IT-9, no. 4, pp. 64–74, Apr. 1963.

[25] I. Jacobs and E. Berlekamp, "A lower bound to the distribution of computation for sequential decoding," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 2, pp. 167–174, Apr. 1967.

[26] D. D. Falconer, "A hybrid coding scheme for discrete memoryless channels," *Bell Syst. Tech. J.*, vol. 48, no. 3, pp. 691–728, Mar. 1969.

[27] F. Jelinek, "An upper bound on moments of sequential decoding effort," *IEEE Trans. Inf. Theory*, vol. IT-15, no. 1, pp. 140–149, Jan. 1969.

[28] T. Hashimoto and S. Arimoto, "Computational moments for sequential decoding of convolutional codes," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 5, pp. 584–591, Sep. 1979.

[29] E. Arikan, "An upper bound on the cutoff rate of sequential decoding," *IEEE Trans. Inf. Theory*, vol. IT-34, no. 1, pp. 55–63, Jan. 1988.

[30] P. Narayan and D. L. Snyder, "Cut-off rate channel design," in *Communications and Cryptography: Two Sides of One Tapestry*, R. E. Blahut, D. J. Costello, Jr., U. Maurer, and T. Mittelholzer, Eds. New York, NY, USA: Springer, 1994, pp. 315–322.

[31] M. S. Pinsker, "On the complexity of decoding," *Problemy Peredachi Informatsii*, vol. 1, no. 1, pp. 113–116, 1965.

[32] J. Massey, "Capacity, cutoff rate, and coding for a direct-detection optical channel," *IEEE Trans. Commun.*, vol. COMM-29, no. 11, pp. 1615–1621, Nov. 1981.

[33] E. Arıkan, "On the origin of polar coding," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 2, pp. 209–223, Feb. 2016.

[34] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051–3073, Jul. 2009.

[35] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 2, pp. 284–287, Mar. 1974.

[36] P. L. McAdam, "MAP bit decoding convolutional codes," Ph.D. thesis, Dept. Elect. Eng., Univ. Southern California, Los Angeles, CA, USA, 1974.

[37] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[38] R. G. Gallager, "The random coding bound is tight for the average code (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 2, pp. 244–246, Mar. 1973.

[39] Y. Domb, R. Zamir, and M. Feder, "The random coding bound is tight for the average linear code or lattice," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 121–130, Jan. 2016.

[40] I. Csiszár, "Linear codes for sources and source networks: Error exponents, universal coding," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 4, pp. 585–592, Jul. 1982.

[41] T. Ho *et al.*, "A random linear network coding approach to multicast," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4413–4430, Oct. 2006.

[42] M. Effros *et al.*, "Linear network codes: A unified framework for source, channel, and network coding," in *Advances in Network Information Theory* (DIMACS Series in Discrete Mathematics and Theoretical Computer Science), vol. 66. Providence, RI, USA: AMS, 2003, pp. 197–216.

[43] J. Wolfowitz, *Coding Theorems of Information Theory*. Berlin, Germany: Springer, 1961.

[44] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

[45] P. M. Ebert, "Error bounds for parallel communication channels," MIT Res. Lab. Electron., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep., 1966.

[46] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 3, pp. 357–359, May 1973.

[47] G. Dueck and J. Körner, "Reliability function of a discrete memoryless channel at rates above capacity (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 1, pp. 82–85, Jan. 1979.

[48] M. P. C. Fossorier and S. Lin, "Soft-decision decoding of linear block codes based on ordered statistics," *IEEE Trans. Inf. Theory*, vol. 41, no. 5, pp. 1379–1396, Sep. 1995.

[49] Y. Kaji and D. Ikegami, "Decoding linear block codes using the ordered-statistics and the MLD techniques," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun./Jul. 2002, p. 144.

[50] J. Hagenauer and L. Papke, "Decoding 'turbo'-codes with the soft output Viterbi algorithm (SOVA)," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun./Jul. 1994, p. 164.

[51] K. Kim, J. W. Choi, A. C. Singer, and K. Kim, "A new adaptive turbo equalizer with soft information classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2010, pp. 3206–3209.

[52] X. Wang and H. V. Poor, "Iterative (turbo) soft interference cancellation and decoding for coded CDMA," *IEEE Trans. Commun.*, vol. 47, no. 7, pp. 1046–1061, Jul. 1999.

[53] S. Song, A. C. Singer, and K.-M. Sung, "Soft input channel estimation for turbo equalization," *IEEE Trans. Signal Process.*, vol. 52, no. 10, pp. 2885–2894, Oct. 2004.

[54] J. B. Anderson, "Limited search trellis decoding of convolutional codes," *IEEE Trans. Inf. Theory*, vol. 35, no. 5, pp. 944–955, Sep. 1989.

[55] G. Foschini, "A reduced state variant of maximum likelihood sequence detection attaining optimum performance for high signal-to-noise ratios," *IEEE Trans. Inf. Theory*, vol. IT-23, no. 5, pp. 605–609, Sep. 1977.

[56] J. Anderson and S. Mohan, "Sequential coding algorithms: A survey and cost analysis," *IEEE Trans. Commun.*, vol. COMM-32, no. 2, pp. 169–176, Feb. 1984.

[57] S. J. Simmons, "Breadth-first trellis decoding with adaptive effort," *IEEE Trans. Commun.*, vol. 38, no. 1, pp. 3–12, Jan. 1990.

[58] S. J. Simmons and P. Senyshyn, "Reduced-search trellis decoding of coded modulations over ISI channels," in *Proc. IEEE GLOBECOM*, Dec. 1990, pp. 393–396.

[59] G. Marino, R. Raheli, G. B. DiDonna, and G. Picchi, "Applications of reduced state sequence estimation to terrestrial digital radio links," in *Proc. IEEE Int. Conf. Commun.*, May 1994, pp. 347–352.

[60] M. V. Eyuboglu and S. U. H. Qureshi, "Reduced-state sequence estimation for coded modulation of intersymbol interference channels," *IEEE J. Sel. Areas Commun.*, vol. 7, no. 6, pp. 989–995, Aug. 1989.
[61] W. Sheen and G. L. Stuber, "Error probability for reduced-state sequence estimation," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 3, pp. 571–578, Apr. 1992.
[62] M. K. Hanawal and R. Sundaresan, "Guessing revisited: A large deviations approach," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 70–78, Jan. 2011.
[63] R. Sundaresan, "Guessing based on length functions," in *Proc. Int. Symp. Inf. Theory*, 2007, pp. 716–719.
[64] M. M. Christiansen, K. R. Duffy, F. du Pin Calmon, and M. Médard, "Guessing a password over a wireless channel (on the effect of noise non-uniformity)," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, 2013, pp. 51–55.
[65] A. Beirami, R. Calderbank, M. Christiansen, K. Duffy, A. Makhdoumi, and M. Médard, "A geometric perspective on guesswork," in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput.*, 2015, pp. 941–948.
[66] M. M. Christiansen, K. R. Duffy, F. du Pin Calmon, and M. Médard, "Brute force searching, the typical set and guesswork," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 1257–1261.
[67] A. Beirami, R. Calderbank, K. Duffy, and M. Médard, "Quantifying computational security subject to source constraints, guesswork and inscrutability," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2015, pp. 2757–2761.
[68] J. T. Lewis, C.-E. Pfister, and W. G. Sullivan, "Entropy, concentration of probability and conditional limit theorems," *Markov Process. Rel. Fields*, vol. 1, no. 3, pp. 319–386, 1995.
[69] J. T. Lewis and C.-E. Pfister, "Thermodynamic probability theory: Some aspects of large deviations," *Russian Math. Surv.*, vol. 50, no. 2, pp. 279–317, 1995.
[70] M. Médard, "A coding theorem for multiple-access decorrelating channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Aug. 1998, p. 215.
[71] X. Chen and D. Leith, "Frames in outdoor 802.11 WLANs provide a hybrid binary-symmetric/packet-erasure channel," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2015, pp. 6128–6132.

Ken R. Duffy is a Professor at Maynooth University where he is currently the Director of the Hamilton Institute. He received the B.A.(mod) and Ph.D. degrees in mathematics from the Trinity College Dublin. His primary research interests are in probability and statistics, and their applications in science and engineering.

Jiange Li received the B.S. degree in applied mathematics from Harbin Institute of Technology in 2009, the M.Sc. and Ph.D. degrees in mathematics from the University of Delaware in 2011 and 2016, respectively. From January 2017 to May 2018, he was a postdoc associate at the Research Laboratory of Electronics (RLE) at the Massachusetts Institute of Technology (MIT). He is now a postdoc fellow at the Einstein Institute of Mathematics at the Hebrew University of Jerusalem. His research interests include probability theory, geometric functional analysis, convex geometry, combinatorics, and information theory.

Muriel Médard is the Cecil H. Green Professor in the Electrical Engineering and Computer Science (EECS) Department at MIT and leads the Network Coding and Reliable Communications Group at the Research Laboratory for Electronics at MIT. She has co-founded three companies to commercialize network coding, CodeOn, Steinwurf and Chocolate Cloud. She has served as editor for many publications of the Institute of Electrical and Electronics Engineers (IEEE), of which she was elected Fellow, and she has served as Editor in Chief of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. She was President of the IEEE Information Theory Society in 2012, and served on its board of governors for eleven years. She has served as technical program committee co-chair of many of the major conferences in information theory, communications and networking. She received the 2009 IEEE Communication Society and Information Theory Society Joint Paper Award, the 2009 William R. Bennett Prize in the Field of Communications Networking, the 2002 IEEE Leon K. Kirchmayer Prize Paper Award, the 2018 ACM SIGCOMM Test of Time Paper Award and several conference paper awards. She was co-winner of the MIT 2004 Harold E. Edgerton Faculty Achievement Award, received the 2013 EECS Graduate Student Association Mentor Award and served as Housemaster for seven years. In 2007 she was named a Gilbreth Lecturer by the U.S. National Academy of Engineering. She received the 2016 IEEE Vehicular Technology James Evans Avant Garde Award, the 2017 Aaron Wyner Distinguished Service Award from the IEEE Information Theory Society and the 2017 IEEE Communications Society Edwin Howard Armstrong Achievement Award. She is a member of U.S. National Academy of Inventors.